தரவுத் தொகுதிமொழியியல்

(Corpus Linguistics)

சங்கரவேலாயுதன் இராசேந்திரன்
அமிர்தா விஷ்வவித்யபீடம், கோயம்புத்தூர்
rajushush@gmail.com

1 அறிமுகவுரை

அனுபவவாத ஆய்வுகள் எந்த எழுதப்பட்ட அல்லது பேசப்பட்ட உரைகளையும் பயன்படுத்தி அதன் மேல் செய்யப்படுவதாகும். இம்மாதிரியான தனிப்பட்ட உரைகள் பலவகையான இலக்கிய மற்றும் மொழியியல் ஆய்வுகளுக்கு அடிப்படையாக அமைகின்றது. எடுத்துக்காட்டாக, ஒரு செய்யுளின் அல்லது ஒரு புதினத்தின் நடையியல் ஆய்வு அல்லது ஒரு தொலைக்காட்சிப் பேச்சின் உரையாடல் ஆய்வு என்பனவற்றைக் கூறலாம். ஆனால் அனுபவவாத மொழியியலின் ஒரு வடிவுக்கு அடிப்படையாக அமையும். ஒரு குறிப்பிட்ட பகுதியைப் பரிசோதிக்கும் வழிகளிலிருந்து பல அடிப்படையான வழிகளில் வேறுபடும். கொள்கை அடிப்படையில் ஒரு உரைக்குக் கூடுதலான எந்தச் சேகரிப்பு தொகுதியும் தரவுத் தொகுதி அல்லது தரவுத்தொகுதி எனப்படும். தரவுத்தொகுதி என்று தமிழில் அழைக்கப்படும். corpus என்பது body என்பதன் இலத்தீன் சொல்லாகும். எனவே தரவுத்தொகுதி என்பதை எந்த ஒரு உரையின் உடல் என்று கூறலாம். ஆனால் தரவுத்தொகுதி என்ற சொல் தற்தால மொழியியல் என்ற சூழலில் பயன்படுத்தப்படும் போது எளிய விளக்க உரையைத் தாண்டி கூடுதல் சிறப்பு அர்த்தங்களைக் கொண்டிருக்கும் போக்கு விரிதரவில் காணப்படும்.

தரவுத்தொகுதியின் இரண்டு முக்கிய பகுதிகள்

1. தகவல் உரைநடையின் உரைகள்
2. கற்பனை உரைநடையின் உரைகள்

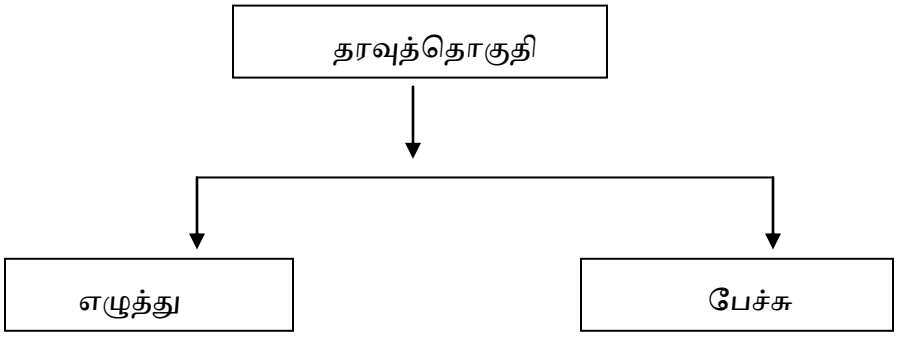
தகவல் உரைகள் பின்வருவனவற்றை உள்ளடக்குகின்றது:

- பத்திரிக்கை : அறிக்கை (அரசியல், விளையாட்டு, சமுதாயம், வட்டாரச் செய்திகள், பொருளாதாரம், பண்பாடு)
- பத்திரிக்கை : இதழாசிரியர் கருத்து (நிறுவனம் சார்ந்தவை, பதிப்பாசிரியருக்கு தனிநபர் கடிதங்கள்)
- பத்திரிக்கை : திறனாய்வு (சினிமா, புத்தகம், இசை, நடனம் போன்றவை)
- மதம் (புத்தகங்கள், கால இதழ்கள், சமயக்கட்டுரைகள்)
- திறமைகளும், பொது போக்குகளும் (புத்தகங்கள் கால இதழ்கள்)
- புகழ் வாய்ந்த மரபுச் செய்திகள் (புத்தகங்கள், கால இதழ்கள்)
- கடிதங்கள், வாழ்க்கை வரலாறு, நினைவுகள்
- பலதரப்பட்டவை (அரசாங்க ஆவணங்கள், அடித்தள அறிக்கைகள் தொழில் நிறுவன அறிக்கைகள், கல்லூரிப் பெயர்ப்பட்டியல் தொழில் நிறுவனம், வீடு)
- கற்றல் மற்றும் அறிவியல் எழுத்துரைகள் (இயற்கை அறிவியல் மருத்துவம், கணினித் தொழில் நுட்பம், பொறியியல்)
- சமூக மற்றும் நடத்தை அறிவியல்கள் (அரசியல் அறிவியல், சட்டம் கல்வி)

மனிதவியலிலிருந்து உரைகள்

கற்பனை உரைநடையின் உரைகள் பின்வருவனவற்றை உள்ளடக்கும்:

- பொதுக் கதைகள் (புதினம், சிறுகதைகள்)
- அறிவியல் கதைகள் (புதினங்கள் , சிறுகதைகள்)
- துணிக்கரசி செயல் பற்றிய மற்றும் மேற்கத்திய கதைகள் (புதினங்கள் , சிறுகதைகள்)
- காதல் கதைகள் (புதினங்கள் , சிறுகதைகள்)
- நகைச்சுவை (புதினங்கள் , சிறுகதைகள்)



2 ஆங்கில மொழிக்கான தரவுத்தொகுதிகள்/தரவுத்தொகுதிகள்

2.1 பிரவுன் தரவுத்தொகுதி

முதல் மின் தரவுத்தொகுதி 1961-ல் பிரவுன் பல்கலைக்கழகத்தில் நெல்சன் பிரான்சிஸ் ஷென்றி குச்சேரா (W.Nelson francies and Henry kuchera) என்பவர்களால் உருவாக்கப்பட்ட பிரவுண் விரிதரவாகும் (Brown corpus (the standard sample of present day American English). இது 1961-ஆம் ஆண்டு நாள் காட்டி ஆண்டு யுஎஸ்ஏ-ல் USA அச்சடிக்கப்பட்ட ஆங்கில உரை நடையின்ன தொடர்ச்சியான பனுவலின் 1014312 சொற்களைக்கொண்டது. இது 2000+ சொற்களைக் கொண்ட 500 மாதிரிகளால் (Sample) பகுக்கப்பட்டள்ளது. ஒவ்வொரு மாதிரியும் ஒரு வாக்கியத்தின் தொடக்கத்தில் தொடங்குகிறது. ஆனால் ஒரு பத்தியின் அல்லது ஒரு பெரிய பகுதியின் தொடக்கத்தில் தொடங்குகின்றது என்பது தேவை இல்லை. ஒவ்வொன்றும் 2000 சொற்களுக்குபின் இறுதியும் முதல் வாக்கியத்தில் முடிகிறது. ஒவ்வொரு மாதிரியும் உரைநடையின் நடைகள் மற்றும் வகைகளின் விரிந்த பரப்பு எல்லையைப் பிரதி நிதித்துவம் செய்யும். செய்யுள்களும், நாடகங்களில் உள்ள உரைகளும் எழுத்துக் கருத்தாடல்களுக்கு அப்பாற்பட்ட பேச்சு கருத்தாடலின் கற்பனையான பொழுது போக்காக இருப்பதன் காரணமாக அவைகள் விலக்கி வைக்கப்படுகிறது. கதைகள் உட்படுத்தப்பட்டுள்ளன. ஆனால் 50 விழுக்காடுக்கு மேல் உரையாடல் உள்ள மாதிரிகள்

எடுத்துக்கொள்ளப்படமாட்டாது. முஹ்திரிகள் அவற்றின் பிரதிகளித்துவப் பின்பிற்கேற்ப வேறு எந்த அகவயமாகத் தீர்மானிக்கப்பட்ட சிறப்புக்காகவும் தேர்ந்தெடுக்கப்படவில்லை.

2.2 லோப் தரவுத்தொகுதி

லங்காஸ்டர் ஒசலோ பெர்கன் தரவுத்தொகுதி (Lancaster Oslo / Bergen (LOB)) என்பது பிரிட்டிஷ்ஷ் ஆங்கிலத்தின் ஒரு மில்லியன் சொல் சேகரிப்பாகும். இது லங்காஸ்டர் பல்கலைக்கழகத்தைச் சார்ந்த ஜியாபிரே லீச் (Geofferey leech) மற்றும் ஒசலோ பல்கலைக் கழகத்தைச் சார்ந்த ஸ்டிக் ஜோகன்சன் (Slig Johansson) என்போர் பெர்கனில் உள்ள மனிதவியலுக்கான நார்வேஜியன் கணிப்பு மைத்துடன் இணைந்து தொகுக்கப்பட்டதாகும். லோப் யுகே–இல் (UK) 1961-ஆம் வருடத்திற்குள் வெளிப்யிடப்பட்ட மூலப்பெருட்களைக் கொண்டது. இது பிரவுன் விரிதரவைப் போல் உரை மாதிரிகளைக் கொண்டது. லோப் தரவுத்தொகுதியின் திட்ட நெறிமுறை யுஎஸ்ஏ இலும், யுகே-விலும் பயன்படுத்தப்படும் ஆங்கிலத்தின் இரு வகைகளுக்கிடையே உள்ள எதிர்கால ஒப்பீட்டிற்கு வேண்டி பிரவுன் தரவுத்தொகுதிடன் ஒற்றுமைக் கொண்டுள்ளது.

2.3 ஆங்கிலத்தில் ஆஸ்டிரேலியன் தரவுத்தொகுதி

ஆங்கிலத்தில் ஆஸ்டிரேலிய தரவுத்தொகுதி (Australian corpus of English (ACE)) என்பது ஆஸ்டிரேலியாவிலுள்ள மாக்குயர் பல்கலைக் கழகத்தில் (Macquaire University) மொழியியல் துறையில் 1986-1990 என்ற கால கட்டத்தில் தொகுக்கப்பட்டது. ஏசியி 1986க்குள் வெளிப்யிடப்பட்டது மூலப் பொருட்களைக் கொண்டிருக்கிறது. ஏசியி பல்வேறுபட்ட மொழியியல் ஆய்வுகளில் உதவுவதற்காகத் திட்டமிடப்பட்ட ஆஸ்டிரேலியாவில் முதலில் தொகுக்கப்பட்ட பலபடித்தான விரிதரவாகும். இது தற்கால ஆஸ்டிரேலிய ஆங்கிலத்தின் திறனான மாதிரியாகும். இது ஆஸ்டிரேலியாவில் கூடுதல் சிறப்புப் பண்புள்ள ஒரு படித்தான தரவுத்தொகுதிகளுடன் ஒப்பீட்டுப் பார்ப்பதற்கு

வேண்டி நோக்கிட்டு விரிதரவாகச் செய்யப்படுகிறன்றது. இது பிரவுன் விரிதரவிலும், லோப் விரிதரவிலும் பிரிதி நிதித்துவம் தரப்பட்டுள்ள இனங்களின் சரிசமத்திற்குப் பொருத்துவத நோக்கமாகக் கொண்டது. எனவே இது ஏறக்குறைய ஒவ்வொரு வகைப்பாட்டிலும் 2000 சொல் மாதிரிகளின் நிகரான குழுமத்தை உருவாக்குகிறது. துணை இனங்களையும், விக்ஷயப்பரப்புகளையும் பொறுத்துவது தேவை என்பதால் தரவுத்தொகுதி வகைப்பட்டிற்குள்ளும் மாதிரிச் செயல் முறைகள் பெரும்பாலும் திட்டமிடப் பட்டுள்ளது. (வரைமுறையற்றதல்ல)

2.4 எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி

எழுதப்பட்ட நியூசிலாந்து ஆங்கிலத்தின் வெல்லிங்டன் தரவுத்தொகுதி (Wellington Written New Zealand English) என்பது 1982 கால கட்டத்தில் வெல்லிங்டன் விக்டோரியா பல்கலைக் கழகத்தில் உருவாக்கப்பட்டது. நியூசிலாந்து ஆங்கில தரவுத்தொகுதியின் நோக்கம், பிரவுன் விரதரவு, லோப் தரவுத்தொகுதி ஏசியி தரவுத்தொகுதி என்பதுடன் நேரடியாக ஒப்பீடு செய்வதை அனுமதிக்கும் எழுத்து நியூசிலாந்து ஆங்கிலத்தின் கணினியாக்கம் செய்யப்பட்ட மாதிரியைத் தருகிறது. முக்கிய உரை வகைப்பாடுகள் எவ்வளவு இயலுமோ அவ்வளவு லோப் தரவுத்தொகுதிடன் பொருந்தும்படி வரிசைப் படுத்தப்பட்டுள்ளது. கற்பனை உரைநடையில் குறிப்பாக கதைகளில் தேவைகருதிச் சில மாற்றங்கள் செய்யப்பட்டுள்ளன.

இந்திய ஆங்கிலத்தில் கோலப்பூர தரவுத்தொகுதி (Kollapur Corpus of India)எஸ்-வி சாஸ்திரியும். அவரைச் சார்ந்தவர்களும் இணைந்து கோலாப்பூரில் உள்ள சிவாஜி பல்கலைக்கழகத்தில் 1988-ஆம் ஆண்டில் வெளிப்யிடப்பட்ட விசயங்களிலிருந்து எடுக்கப்பட்ட இந்திய ஆங்கிலத்தின் ஒப்பீட்டு ஆய்வுகளுக்கு வேண்டிய மூல சாதனங்களாகப் பயன்படுகின்றது. இது இந்திய ஆங்கிலத்தின் விரிந்த வர்ணனைக்குக் கொண்டுச் செல்லும் என்று எதிர்பார்க்கப்படுகிறது. லோம் மற்றும் பிரவுன் தரவுத்தொகுதிகளை உருவாக்க பயன்படுத்த திட்டத்தைப் பின்பற்றி கோலாப்பூர்

தரவுத்தொகுதிக்கு 15 விக்ஷய வகைப்பாடுகளிலிருந்து உரைகள் தேர்ந்தெடுக்கப்பட்டுள்ளன.

2.5 லோப் தரவுத்தொகுதி

ஃப்லோப் (ஃப்ரெல்பர்க் லோப் தரவுத்தொகுதி (Freiburg lob Corpus for British English (Flob)) மற்றும் லோப் தரவுத்தொகுதிகளுடன் பொருத்தும் ஒரு குழும தரவுத்தொகுதிகளைத் தொகுக்கும் முயற்சியின் விளைவாகும். இது 1990-களின் தொடக்க நிலை மொழியைப் பிரதிநிதித்துவம் செய்யும் என்ற நிலையில் வேறுபடும் இத்திடம் 1990-ஏப்ரலில் ஜெர்மனியில் கிருஷ்டியன் மேரியின் கீழ் தொடங்கப்பட்டது.

இது மொழியாலாளர்களுக்குப் பின் வரும் செயல்களில் உதவுகிறது.

- இன்றைய ஆங்கிலத்தில் மொழிசார் மாற்றத்தின் மீதான இன்றைய கருதுகோளாப் பரிசோதிக்க.
- சொல் நிகழ்வெண்களின் முறையான ஒப்பீடுகளின் வழி இலக்கியத்தில் இதுவரை கண்டுப்பிடிக்கப்படாத மாற்றங்களைக் கண்டுபிடிக்க
- ஒரு கால இட வழக்குகளின் (British and America) தொடர்ச்சியான மாற்றம் மற்றும் நடையியல் வேறுபாடு இவற்றிற்கு இடையிலான சார்பையும் மெய்யான இருகால வளர்ச்சியையும் ஆயும் முக்கியமான நெறிமுறைச் சிக்கல்களில் ஒன்றை முறையாகக் கையாளுவது. இப்புதிய பிரிட்டிக்ஷ் மற்றும் அமெரிக்க தரவுத்தொகுதிகளின் கூடுதல் பயன் என்னவென்றால் அவை இந்திய, ஆஸ்திரேலிய மற்றும் நியூசிலாந்து தரவுத்தொகுதிகளுடன் (பிந்தைய 1980-களில் உளடள மொழி பயன்பாட்டை பிரதிநிதித்துவம் செய்யும் மாதிரிகள்) ஒப்பிடமூல லோப் மற்றும் பிரவுன் தரவுத்தொகுதிகளைவிடக் கூடுதல் பொருத்தமான தரவு மையங்களைத் தருகின்றன.

2.6 பிரிட்டிஷ் தேசிய தரவுத்தொகுதி

பிரிட்டிஷ் தேசிய தரவுத்தொகுதி (British National Corpus (BNC) என்பது மூலங்களின் விரிந்த பரப்பெல்லையிலிருந்து பெறப்பட்டு எழுதப்பட்ட மற்றும் (10%) பல வயதினர்கள், வட்டாரங்கள், சமூக வகுப்புகள் இவற்றிலிருந்து பெறப்பட்ட எழுதப்படாத முறைசாரா உரையாடல்கள் மற்றும் முறையான வணிக மற்றும் அரசாங்கக் கூட்டங்களிலிருந்து வானொலி மற்றும் தொலைபேசி வரையிலான பரப்பெல்லையைக் கொண்ட சூழல்களின் எல்லா வகைகளிலிருந்தும் சேகரிக்கப்பட்ட பேச்சு மாதிரிகள் என்பனவற்றை உள்ளடக்கும்.

இப்பொதுத்தன்மை பிஎன்சி விரிதரவை அகராதியியல், பொருண்மையியல், உருபனியல், மொழிப் பகுப்பாய்வு, செயற்கை அறிவுநுட்பம், பேச்சு அறிதல் மற்றும் கூட்டிணைப்பு, இலக்கிய ஆய்வுகள் மற்றும் முக்கிய மொழியியலின் எல்லாப் பரப்புகளையும் உள்ளடக்கும் ஆய்வு நோக்கங்களின் விரிந்த வகைக்குப் பயனுள்ளதாகச் செய்கின்றது.

பிஎன்சி பேச்சு உரையாடல் மற்றும் ஒருவர் உரையாடல் இவற்றிலிந்து ஒலிபெயர்ப்பு செய்யப்பட்ட 863 உரைகளை உள்ளடக்கிய 4124 உரைகளைக் கொண்ட 100106008 சொற்களை உள்ளடக்கியது. ஒவ்வொரு உரையும் மேலும் எழுத்துவடிவ வாக்கிய அலகுகளாகப் பிரிக்கப்பட்டுள்ளது. இதற்குள் ஒவ்வொரு சொல்லுக்கும் சொல் வகுப்பு (சொல்வகைப்பாடு) குறியம் (code) தரப்பட்டுள்ளது. பிரித்தலும் சொல் வகைப்படுத்தலும் (claws) சொல்வகைப்பாட்டு அடையாளப்படுத்தியால் தானியக்கமாகச் செய்யப்படுகின்றது. பிஎன்சி விரிதரவில் பயன்படுத்தப்படும் வகைப்பாட்டுத் திட்டம் 65 சொல் வகைப்படுகளை வேறுபாடுகளை வேறுபடுத்துகின்றது.

2.7 அமெரிக்க தேசிய தரவுத்தொகுதி

அமெரிக்க தேசிய தரவுத்தொகுதி (American National Corpus (ANC)) பிஎன்ச தரவுத்தொகுதி ஒப்பிடத்தக்க அமெரிக்க ஆங்கிலத்தை உட்படுத்திய தரவுத்தொகுதியின் உருவாக்கத்திற்கு உதவியது. இதன் நோக்கம் பிஎன்சி தரவுத்தொகுதிப் பொதுவினங்களுடன் ஒப்பிடத்தக்கக் குறைந்தது 100 மில்லியன் சொற்களைப் பெறுவதாகும். இது மொழித் தொழில் நுட்ப ஆய்வுக்கு உதவுவதுடன் எல்லா மட்டங்களிலும் மொழிப் பகுத்தாய்வுக்கும் கல்விக்கும் பயன்படவேண்டி செழுமையான தேசிய மூலத்தைத் தருகின்றது. எஎன்சி அமெரிக்க ஆங்கில அகராதிகளின் வெளிப்யீட்டார்கள் மற்றும் மொழி ஆய்வில் ஆர்வமுள்ள நிறுவனங்கள் இவற்றின் கூட்டிணைப்பின் (consortium) பங்களிப்பு வழி உருவாக்கப்பட்டது. மொழித் தரவுக் கூட்டிணைப்பும் (linguistics Data consortium LDC) சாதன மற்றும் பொருளாதார ஆதரவும் தந்தது. ஏஎன்சி தரவுத்தொகுதியின் 10 மில்லியன் சொற்களின் முதல் வெளிப்யீட்டில் உள்ளடக்கப்பட்ட உரைகள் முதலில் பெறப்பட்டவையாகும். எனவே, தரவுத்தொகுதி சரிநிகரானதல்ல பேச்சுப் பகுதி 3224388 சொற்களையும் எழுத்துப்பகுதி 8283828 சொற்களையும் கொண்டுள்ளது.

2.8 ஆங்கில வங்கி

ஆங்கில மொழி வங்கி (Bank of English) சொற்கள், இலக்கணம் மற்றும் பயன்பாடு (usage) என்பவனவற்றின் பகுப்பாவிற்கு வேண்டி வடிவமைக்கப்பட்ட தற்கால ஆங்கில மொழியின் மாதிரிகளின் சேகரிப்பாகும். ஆங்கில மொழி வங்கி 1991-ன் (Cobuild (Harper Collins Publishers))மற்றும் யுகேயில் உள்ள பிர்மிங்காம் பல்கலைக்கழகம் இவற்றால் 1991-ல் வெளிப்யிடப்பட்டது. இது புதிய சாதனங்களின் நிரந்தர சேர்ப்பினால் தொடர்ந்து வளர்கின்றது. சேகரிப்பு எழுதப்பட்ட மற்றும் பேச்சின் வேறுபட்ட வகைகளின் விரிந்த பரப்பெல்லையைக் கொண்டது. இது நூற்றுக்கணக்கான வேறுபட்ட மூலங்களிலிருந்து ஆங்கில மொழியின் மாதிரிகளைக் கொண்டிருக்கின்றது. எழுதப்பட்ட உரைகள் செய்தித்தாள்கள், இதழ்கள், (magazines) கதைப் புத்தகங்கள் மற்றும் கதையல்லாப் புத்தகங்கள், சிற்றேடுகள் (brpcjers) துண்டுப்பத்திரிக்கைகள் (leaflets) அறிக்கைகள்,

கடிதங்கள் மற்றும் பிறவற்றிலிருந்து எடுக்கப்பட்டவைகளாகும். பேச்சுச் சொற்கள் அன்றாட சாதாரண உரையாடல்கள், வானொலி ஒலிபரப்புகள், கூட்டங்கள் நேர்காணல்கள் மற்றும் விவாதங்கள் போன்றவற்றால் பிரதிநிதித்துவம் செய்யப்பட்டுள்ளன.

3 இந்திய மொழிகளுக்கான தரவுத் தொகுதிகள்

3.1 இந்திய மொழிகளின் MIT தரவுத்தொகுதி

இந்திய மொழிகளுக்கு தரவுத்தொகுதி உருவாக்குவது 1991-ல் தொடங்கப்பட்டது. இது இந்திய அரசின் மின்னுத் துறை எல்லா இந்திய மொழிகளுக்கும் இயந்திரத்தால் படிக்கவியலும் உரைகளின் தரவுத்தொகுதிகளஇ உருவாக்க வேண்டி இந்திய மொழிகளுக்குத் தொழில் நுட்ப வளர்ச்சி (Technological Development for Indian Languages (TDIL)) என்ற திட்டத்தைத் தொடங்கியது. இதே நேரத்தில் மொழி ஆய்வுக்கும் (சொல்வகைப்பாடு அடையாளப்படுத்தி, உரை குறியாக்கி, புள்ளியல் எண்ணி, எழுத்துத் திருத்தி, உருபனியல் பகுப்பாய்வி போன்றவற்றை உருவாக்க) ஆங்கிலத்திலிருந்து இந்திய மொழிகளுக்கு எந்திர உதவிசார் மொழிபெயர்ப்புக்கு (machine-aided translation) வேண்டி கருவிகளை வடிவமைத்தல் என்ற செயல்பாடுகளுக்கு மென்பொருள்களை உருவாக்கவும் ஊக்கம் தரப்பட்டது. டெல்கியில் உள்ள இந்தியத தொழில் நுட்ப நிறுவனம் (Indian institute of technology) இந்திய ஆங்கிலம் , இந்தி மற்றும் பஞ்சாபி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் தமிழ், தெலுங்கு, கன்னடா மற்றும் மலையாள மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. பூனேயில் உள் டெக்கான் கல்லூரி மராத்தி மற்றும் குஜராத்தி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. புபனேஸ்வரிலுள்ள பயன்பாட்டு மொழி அறிவியல்களின் இந்திய நிறுவனம் ஒரியா, வங்காளம், அசாமீஸ் ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது. வாரணாசியிலுள்ள சம்பூர்ணநந்தா சமஸ்கிருதப் பல்கலைக்கழகம் சமஸ்கிருத மொழிக்கு

தரவுத்தொகுதி உருவாக்கியது. அலிகார் முஸ்லீம் பல்கலைக்கழகம் உருது, சிந்தி மற்றும் காஷ்மீரி ஆகிய மொழிகளுக்கு தரவுத்தொகுதிகளை உருவாக்கியது.

கான்பூரிலுள்ள இந்தியத் தொழில் நுட்ப நிறுவனம் மொழி பகுத்தாய்விற்கும் இயந்திர மொழிபெயர்ப்புக்கும் கருவிகளையும் மென்பொருள்களையும் வடிவமைக்கும் பொறுப்பை ஏற்றது. மைசூரிலுள்ள இந்திய மொழிகளின் மைய நிறுவனம் எல்லா மொழிகளின் முழு தரவுத்தொகுதி தரவுமையத்தைச் சேகரித்துப் பாதுகாக்கும் பொறுப்பை ஏற்றது. 1995-ல் இத்திட்டத்தின் முடிவுற்றபோது மின்னனு நிறுவனம் இது நீண்டகால ஆய்வையும் பெரிய முதலீட்டையும் வேண்டும் என்று உணர்ந்து திட்டத்தைத் தொடராமல் நிறுத்தியது. இருப்பினும் தற்போது இந்திய அரசின் MIT புதிய வெளிப்ச்சத்திலும் பார்வையிலும் இச்செயல்பாட்டை மீண்டும் நடைமுறைப்படுத்தியுள்ளது.

3.2 இந்திய மொழிகளுக்கான தரவுத்தொகுதி

இந்திய மொழிகளுக்கான தரவுத்தொகுதி பற்றி Jayaram மற்றும் Rajashree விரிவாக விளக்கியுள்ளனர். மேலும் இந்திய மொழிகளின் நடுவண் நிறுவனம் இந்தி, பஞ்சாபி, காஷ்மீரி, உருது, கன்னடம், தமிழ், தெலுங்கு, மலையாளம், ஒரியா, அசாமி மற்றும் பெங்காலி போன்ற பன்னிரெண்டு மொழிகளுக்கும் கிட்டத்தட்ட மூன்று மில்லியன் சொற்கள் வருமாறு விரிதரவை உருவாக்கியுள்ளது.

தரவுத்தொகுதியின் நோக்கம்

தரவுத்தொகுதி இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட தரவுத்தொகுதிகள் பல்நோக்கு தரவுத்தொகுதிகள் ஆகும். இந்த ஒரு விரிதரவை உருவாக்க செலவான தொகை மற்றும் தொகுக்க தேவையான நேரம் என்பன பல்வேறுபட்ட மொழியியல் ஆய்வுகளுக்கு இவற்றை அடிப்படை மூலவளமாக பயன்படுத்த ஊக்கம்

அளித்து வருகிறது. ஒருவர் தரவுத்தொகுதி செய்ய எல்லா ஆய்வுகளுக்கும் மிகப் பொருத்தமான முன்மாதிரி என்று எடுத்துக்கொள்ள இயலாது. தரவுத்தொகுதியின் நோக்கம் பின்வருமாறு:

1. வேறுபட்ட உரை வகைகளிலிருந்து சொற்கள் மற்றும் வாக்கியங்களின் மாதிரி.
2. வேறுபட்ட உரை வகைகளில் சொற்களின் நிகழ்வெண்
3. இலக்கண வகைப்பாடுகள் அடிப்படையில் சொற்களின் பட்டியல்
4. வேறுபட்ட சுழல்களில் சொற்களின் தேர்வு
5. தனிநிலை சொற்கள் சேர்ந்து வருகை மற்றும் இலக்கண – தொடரியல் அமைப்பொழுங்கு
6. இலக்கண வகைப்பாடுகள் அடிப்படையில் சொற்களின் பட்டியல்
7. பயன்பாட்டடின் இயற்கையான எடுத்துக்காட்டுகளின் தேர்வுக்குக் கச்சாப்பொருள்.

3.3 EMILLE – தரவுத்தொகுதி

EMILLE [Enabling Minority Language Engineering] தரவுத்தொகுதிகள், Emille திட்டம், யு.கே.யில் உள்ள லங்காஸ்டர் பல்கலைக்கழகம், மைசூரில் உள்ள இந்திய மைய நிறுவனம் இவற்றின் ஒத்துழைப்பால் உருவாக்கப்பட்டது. Emille ஐரோப்பிய மொழி மூலவளங்கள் சங்கத்தால் [European Language Resources Association] வழங்கப்பட்டது. இந்த தரவுத்தொகுதி மூன்று பகுதிகளைக் கொண்டது.

1. ஒரு மொழி தரவுத்தொகுதி [Monologual Corpus]
2. இணை தரவுத்தொகுதி [Parallel Corpus]
3. அடையாளப்படுத்தப்பட்ட தரவுத்தொகுதி [Annotated Corpus]

ஆசாமி, பெங்காளி, குஜராத்தி, இந்தி, கன்னடம், காக்ஷ்மிரி, மலையாளம், மராத்தி, ஒரியா, பஞ்சாபி, சிங்களம், தமிழ், தெலுங்கு, உருது ஆகிய பதினாங்கு தெற்காசிய மொழிகளின் எழுத்து மற்றும் பேச்சுத் தரவை உள்ளடக்கிய ஒரு மொழி தரவுத்தொகுதிகள் உள்ளன. நுஅடைடந ஒரு மொழி தரவுத்தொகுதி கிட்டத்தட்ட 92,799,000 சொற்களை உள்ளடக்கியது. (இது பெங்காளி, குஜராத்தி, இந்தி, பஞ்சாபி, உருது என்ற மொழிகளின் எழுத்துப்பெயர்க்கப்பட்ட பேச்சுத்தரவின் 2,627,000 சொற்களை உள்ளடக்கியுள்ளது) இணை தரவுத்தொகுதிகள் ஆங்கில உரையின் 2,00,000 சொற்களையும், இந்தி, பெங்காளி, பஞ்சாபி, குஜராத்தி, உருது இவற்றின் மொழிபெயர்ப்புகளையும் கொண்டது. அடையாளப்படுத்தப்பட்ட பகுதி உருது மொழியின் சொல்வகைப்பாட்டு அடையாளப்படுத்தப்பட்ட ஒரு மொழிய மற்றும் இணை தரவுத்தொகுதிகளை உள்ளடக்கியது. இத்துடன் வரையறை பயன்பாட்டின் இயல்பைக்காட்ட இருபது இந்தி தரவுத்தொகுதிகளின் கோப்புகளைக் கொண்டுள்ளது. இந்த விரிதரவானது CES Compliant SGML- ஐப் பயன்படுத்திக் குறிக்கப்பட்டுள்ளது. மேலும் Unicode- ஐப் பயன்படுத்தி குறியாக்கம் (Encode) செய்யப்பட்டுள்ளது.

3.4 EMILLE மற்றும் CIIL ஒரு மொழி எழுத்து தரவுத்தொகுதி

எழுத்து தரவுத்தொகுதி டெல்லியில் உள்ள இந்திய தொழில் நுட்ப நிறுவனம் (CIIT),புவனேசுவரத்தின உள்ள பயன்பாட்டு மொழி அறிவியல் நிறுவனம் (ILAS), அலிகாரில் உள்ள அலிகார் முஸ்லீம் பல்கலைக்கழகம் இவற்றுடன் ஒருங்கிணைந்து இந்திய மொழிகளின் மைய நிறுவனத்தால் உருவாக்கப்பட்டது CIIL தரவுத்தொகுதி ஆகும்.

இந்த தரவுத்தொகுதி தொடக்கத்தில் ISCII உரையாக குறியாக்கம் செய்யப்பட்டுள்ளது. இப்பொழுது CES – Compliant SGML குறியீட்டுடன் Unicode – CIIL ஒரு மொழிய தரவுத்தொகுதி 93,530,000 சொற்களைக் கொண்டுள்ளது.

3.5 தமிழுக்கான தரவுத்தொகுதி தயாரித்தல்

தமிழுக்கான தரவுத்தொகுதி தயாரித்தலில் முன்னோடியாக உள்ள நிறுவனம் மைசூரில் உள்ள இந்திய மொழிகளின் மைய நிறுவனம் ஆகும். பல பிற நிறுவனங்களும் தாங்கள் எடுத்துக் கொண்ட இயற்கை மொழியாய்வு திட்டங்களுக்கு ஏற்றவாறு தரவுத்தொகுதிகளைச் சேகரித்துப் பயன்படுத்துகின்றன. மின் வடிவில் உள்ள புத்தகங்களையும் கால இதழ்களையும் நாள் இதழ்களையும் தேவைக்கேற்றவாறு சேகரித்து விரிதரவாகப் பயன்படுத்த இயலும். இந்திய மொழிகளின் மைய நிறுவனம் இந்திய மொழிகளுக்காக தரவுத்தொகுதி தயாரிக்கும் முயற்சியில் தமிழுக்கு விரிதரவைத் தயாரித்து தந்துள்ளது.

தமிழுக்கான தரவுத்தொகுதிப் பொறுத்த வரையில் உரைகளானது தினகரன் பத்திரிக்கையின் இணையத்தளத்திலிருந்து எடுக்கப்பட்டுள்ளது. இந்த தரவுத்தொகுதி இந்திய மொழியின் மைய நிறுவனத்தால் உருவாக்கப்பட்ட தொடக்கால தரவுத்தொகுதியும் உட்படுத்தும். தமிழ் எழுத்து தரவுத்தொகுதியிலிருந்து எடுக்கப்பட்ட சொற்களின் பட்டியல் கீழே கொடுக்கப்பட்டுள்ளது.

1.சினிமா (cinema) திரைப்படம் மற்றும் அது தொடர்பாக செய்திகள் (கிட்டத்தட்ட 1,050,000 சொற்கள்)

2.செய்திகள் (News) கதைச் செய்திகள் (கிட்டத்தட்ட 8,610,000 சொற்கள்)

3.மற்றவை (Others) இணைய தளத்திலிருந்து எடுக்கப்பட்ட பிற கதைகள் (கிட்டத்தட்ட 730,000 சொற்கள்)

4.அரசியல் (Political) அரசியல் செய்தி அறிக்கை (கிட்டத்தட்ட 5,050,000 சொற்கள்)

5.விளையாட்டுகள் (Sports) விளையாட்டுச் செய்தி அறிக்கை (கிட்டத்தட்ட 1,170,000 சொற்கள்)

பிற நிறுவனங்களின் தமிழ் தரவுத்தொகுதிகள்

இயற்கை மொழி ஆய்வில் குறிப்பாக இயந்திர மொழிபெயர்ப்பில் சூடுப்பட்டுள்ள நிறுனங்கள் தமிழுக்கென்று தரவுத்தொகுதிகள் தயாரித்து அவற்றை சொல் வகைப்பாடுகளுக்கு அடையாளப்படுத்தி பயன்படுத்தி வருகின்றன. குறிப்பாகத் தமிழ்ப்பல்கலைக்கழகம், அண்ணாமலைப் பல்கலைக்கழகம், AUKBC ஆராய்ச்சி நிறுவனம், அமிர்தா தொழில் நுட்ப நிறுவனம் என்பன இயந்திர மொழிபெயர்ப்பு ஆய்வில் சூடுபட்டு அதற்கேற்ற தரவுத்தொகுதிகளை உருவாக்கி பயன்படுத்துகின்றன.

4 உரை தரவுத்தொகுதி ஆய்வு

பல மொழிகளில் பெரிய மின் தரவுத்தொகுதிகள் சேகரிக்கப்பட்டபின் அவற்றை ஆய்வு உத்திகளுக்கு உள்ளாக்கும் தேவைவரும். விரிதரவிலிருந்து மொழித்தரவை அணுகுவதற்கும் தேவையான தகவல்களை மீளப்பெறுவதற்கும் மக்கள் ஒழுங்குமுறைகளும் உபாயங்களும் உருவாக்கியுள்ளனர். இவ்வாய்வுக் கருவிகள் பயனுள்ளதாக அமைந்துள்ளன. பலவிதமான ஆய்வு உத்திகள் இருக்கின்றன. எடுத்துக்காட்டாக.

1. புள்ளியல் ஆய்வி (Statistical analyzer)
2. தொடரடைவு ஆய்வி (Concordencer)
3. சொல் சேர்ந்து வருகை ஆய்வி (Lexical collocator)
4. முக்கியச்சொல் காணும் ஆய்வி (Key word finder)
5. அகராதிச்சொல் ஆய்வி (lemmantizer)
6. உருபனியல் ஆய்வியும் உருவாக்கியும் (Marphological analyzer and gernerator)
7. சொல் ஆய்வி (word processer)
8. சொல் வகைப்பாடு அடையாளப்படுத்தி (part of speech tagger)

9. தரவுத்தொகுதி குறிப்புரைப்பான் (corpus annotator)

10. பகுத்துக் குறிப்பான் (Parser)

ஆங்கிலம், பிரஞ்சு, ஜெர்மன் போன்ற மொழிகளுக்கு விரிதரவை ஆய பல மென்பொருள்கள் உருவாக்கப்பட்டுள்ளன. ஆனால் இந்தி, தமிழ் போன்ற இந்திய மொழிகளுக்கு அத்தகைய மென்பொருள்கள் இனிமே தான் உருவாக்கப்பட வேண்டும். இதற்கான முயற்சிகள் எடுக்கப்பட்டு வருகின்றன.

4.1 நிகழ்வெண் ஆய்வு

மொழியில் நீண்ட காலமாகவே புள்ளியியலுடனும் கணக்கியலுடனும் தொடர்புடையது. மொழியியல், நடை அளவியல் (stylometrics) போன்ற இயற்கைமொழி விரிதரவிலிருந்து பெறப்படும் வேறுப்பட்ட புள்ளியியல் மற்றும் அளவில் தகவல்களை வேண்டுகிறது. மொழியின் அட்டவணைப் படுத்தப்படும். உரையிலுள்ள பல முக்கியமான மொழி அமைப்பொழுங்குகளை அணுக உதவுவதால் இது மிக முக்கியமானதாகும். இது உள்ளுணர்வால் அறிய இயலாத தகவல்களைக் தருகின்றது.

4.2 சொல் வருகை ஆய்வி

சொற்களின் சேர்ந்து வருகை உரைகளில் சொற்களின் பங்களிப்பையும் இடத்தையும் புரிந்துக்கொள்ளப் பெரிதும் உதவும் இவ்வாய்வு சொற்களில் எந்த இணைகளுக்கிடையில் கூடுதல் சேர்ந்து வருகை உறவு இருக்கின்றது என்பதை அறிய இயலும். மேலும் இரண்டு சொற்கள் சேர்ந்து வரும் நிகழ்வுத் தகைமையைக் தற்செயலாக விளைவின் அடிப்படையிலான நிகழ்வு தகைமையுடன் ஒப்பிடும் ஒவ்வொரு சொல் இணைக்கும் ஒரு மதிப்பெண் தரப்படும் கூடுதல் மதிப்பெண் சேர்ந்து வருகையைக்காட்டும். இது அகராதியிலும் தொழில் நுட்ப மொழி பெயர்ப்பிலும் பயன்படுத்த விரிதரவிலிருந்து பல்சொல் அலகுகளைப் பிரித்தெடுக்க உதவும். இயந்திர

மொழிபெயர்ப்புக்கும் பொருண்மை வேறுபாட்டிற்கும் இத்தகைய தகவல்களை கைக்கொடுக்கும் இது அர்த்த வேறுபாடுகளை அடையாளம் காண ஒற்றுமையான சொற்களைக் குழும உதவுகிறது.

எ.கா.

மாடிப் - படி Step

வாசற் - படி Step

புத்தகம் படி Read

சொன்னப்படி Manner

ஐந்து லிட்டர் Measure

ஒரே பொருண்மையைக் கொண்ட சொல்லுகளுக்கு இடையே உள்ள பயன்பாட்டு வேறுபாடுகளை வேறுபடுத்த உதவும்.

4.3 சூழலில் முக்கியச் சொல்

தரவுத்தொகுதி ஆய்வில் “KWIC” (Key Word in Context (KWIC))’’ என்ற சொல் அடிக்கடி பயன்படுத்தப்படுகிறது. இது குறிப்பிட்ட சொற்களின் நேர்வுகளை அறிய உதவுகின்றது. ஆய்வுக்கு உட்படும் சொல் ஒவ்வொரு வரியின் மத்தியில் கூடுதலான இடை வெளிப்யுடன் தோன்றும் வேறுபட்ட நோக்கங்களுக்கு வேண்டி சூழலின் நீட்சி குறிப்பிடப்பட்டிருக்கும். இது மையத்திலுள்ள சொல்லின் இரண்டு பக்கங்களிலும் இரண்டு மூன்று அல்லது நான்கு சொற்களின் சூழலைக் காட்டும். இதை N-Gram (bi-gram, trigram, tetra-gram set) என்பர் இவ்வமைப்பொழுங்கு ஒருவரின் தேவை அடிப்படையில் மாறும். சொற்கள், தொடர்கள், எச்சத் தொடர்கள் இவற்றின் ஆய்வின் போது கூடுதலான சூழல் அவற்றைப்புரிந்துக்கொள்ளத் தேவைப்படும். மறஇஉ என்பதை ஒரு வரையாக எண்ணுவது நல்லது. சொற்களின் நிகழ்வெண்ணை மையச்சொல்லின் சூழலில் பரிசோதிக்கவேண்டும். எல்லாத் தகவல்களுகம் எப்போதும்

தேவைப்படுவதில்லை. ஆனால் நாம் தேவைப்படும் போது தகவலைப் பயன்படுத்துகின்றோம்.

KWIC சொல்லால் பொருந்தலை அணுகியபின் நாம் மொழியியல் வர்ணணையில் பல்வேறுபட்ட நோக்கங்களை உருவாக்கலாம் மற்றும் இந்நோக்கங்களை நிறைவேற்ற வழிமுறைகளை வகுக்கலாம். சூழலில் முக்கியத்துவம், சேர்ந்து வரும் சொற்களின் பங்களிப்பு சூழல்களில் சொற்களின் உண்மை நடத்தை, நேர்வின் உண்மைச் சூழ்நிலை மற்றும் சூழல் அடிப்படையிலான கட்டுப்பாடுகள் இவற்றைப் புரிந்துகொள்ள KWIC உதவும்.

குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் உரைகளில் சொற்கள் பயன்படுத்தப்படும் அமைப்பொழுங்கைப் பற்றிய தகவல்களை நமக்குத் தரும் மற்றொரு வகையான ஆய்வாகும். இது எங்கு சொல் நிரல் வாக்கியங்களின் பொருண்மைச் சுமையைத் தீர்மானிக்கின்றதோ அங்கும், எங்கு ஒரு உறுப்பின் தனிப்பட்ட பொருண்மைச் சுமை மற்றொரு உறுப்பின் வருகையால் பாதிக்கப்படுகிறதோ அங்கும் முக்கியத்துவம் வாய்ந்ததாகும். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதல் (தொடர்கள் மற்றும் வாக்கியங்களின்) பகுத்துக் குறித்தலின் போது உறுப்புகளின் செயல்பாட்டு நடத்தையை நேரிட தகவலைத் தருகிறது.

இது பயன்பாட்டில் வழக்கமற்ற ஆனால் தற்சுட்டு வடிவுகளுடன் தனிப்பட்ட உறவு கொண்ட வினைவடிவுகளின் எடுத்துக்காட்டாக (Amuse onself, please one self, lend on self remind one self)வகைமுறையைத் தீர்மானிக்கின்றது. இம்மாதிரியான அமைப்பொழுங்கின் அறிவு, மொழி கற்பவர்களுக்கு இடைப்பட்ட நிலையிலிருந்து உயர்ந்த நிலைக்குச் செல்ல குழுமுதலைப் பயன்படுத்தி நாம் தமிழில் முற்றும் வினைகள், பெரும்பாலான எச்ச வினைகள் கூடுதலாகத் தொடரும் என்றும் பெயர்கள் ஒட்டுக்களாலும், பின்னுருபுகளாலும் தொடரப்படும் என்றும் அறிந்து கொள்ளலாம். குறிப்பிட்ட இடம் சார்ந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல்கள் சொற்களை

வினைக் குழுக்களாகவும் பெயர்க் குழுக்களாகவும் ஆய உதவும். குறிப்பிட்ட இடம் சாரந்த சொற்களைக் குழுமுதலால் கிடைக்கும் தகவல் சொற்களின் குறிப்பிட்ட இடம் சார்ந்த சேர்க்கையின் காரணமாக வரும் சொல்மயக்கத்தைத் தீர்க்கின்றது. பொருண்மையின் நுண்மையான வேறுபாடுகள் சூழல்களின் சொற்களின் வருகை முறையுடன் கூடிய உறுப்புகளுக்கிடையே உள்ள அக உறுப்புக்காளல் பெரும்பாலும் தரப்படுகின்றது. கூட்டுப் பெயர்களுக்கும், கூட்டு வினைகளுக்கும் சொற்களின் குறிப்பிட்ட சேர்க்கையால் உணர்த்தப்படும் பொருண்மைத் தனிப்பட்ட சொற்களின் பொருண்மையிலிருந்து பெற இயலாது.

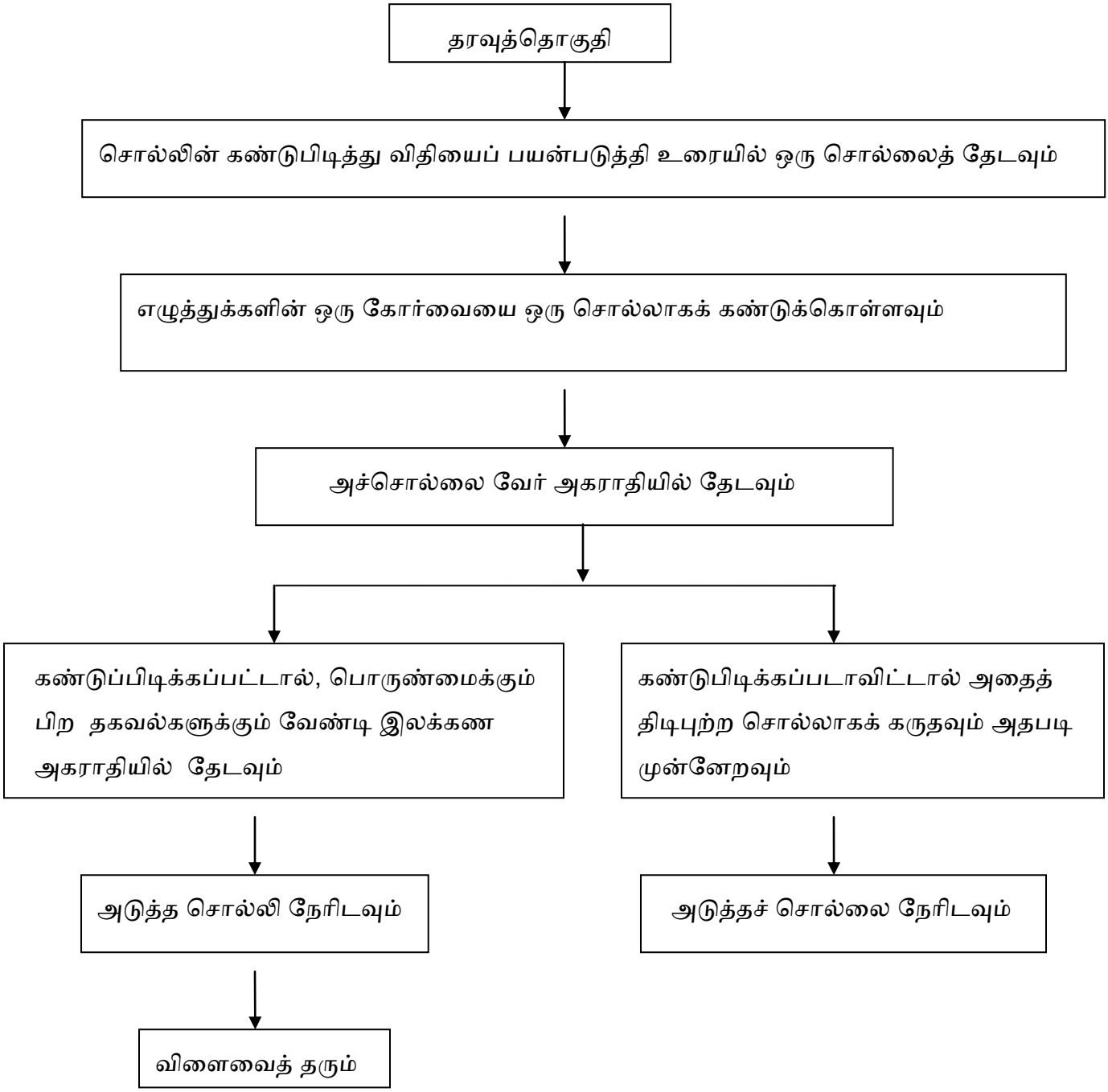
4.4 சொல் பகுப்பாய்வு

சொல் பகுப்பாய்வு (Word Processing) விரிதரவில் பயன்படுத்தப்படும் சொற்களைத் தானியக்கமாக ஆய்வதை உள்ளடக்கும். இதன் முக்கிய நோக்கம், ஒரு உரையின் பகுதியிலிருந்து ஒரு சொல்லைக் கண்டுபிடித்து அதை அதன் பயன்பாட்டின் சூழலிருந்து பிரித்து, அதன் உருபொலியனியல் அமைப்பை ஆய்ந்து, அதன் மூலப் பொருண்மையைப் பெற்று, உரையில் அதன் தொடரியல் பங்களிப்பை விளக்குவது அவசியம்.

சொல் பகுப்பாய்விலிருந்து கிடைக்கும் தகவல் சொல் அர்த்த மயக்கத்தைத் தீர்ப்பதற்கும் (Word Sense Disambiguation (WSD))அகராதி உருவாக்குவதற்கும் பகுத்துக் குறிப்பதற்கும் மொழி கற்பத்றகும் மற்றும் இது போன்ற பலற்றிற்கும் பயனுள்ளதாகும். ஆங்கிலம் போன்ற மொழிகளுக்குப் பல சொல் பகுப்பாய்வுகள் உருவாக்கப்பட்டுள்ளன. தமிழ் போன்ற இந்திய மொழிகளுக்கும் இத்ததகைய பகுப்பாய்வு உருவாக்கப்பட்டு பயன்படுத்ததப்பட்டு வருகின்றன.

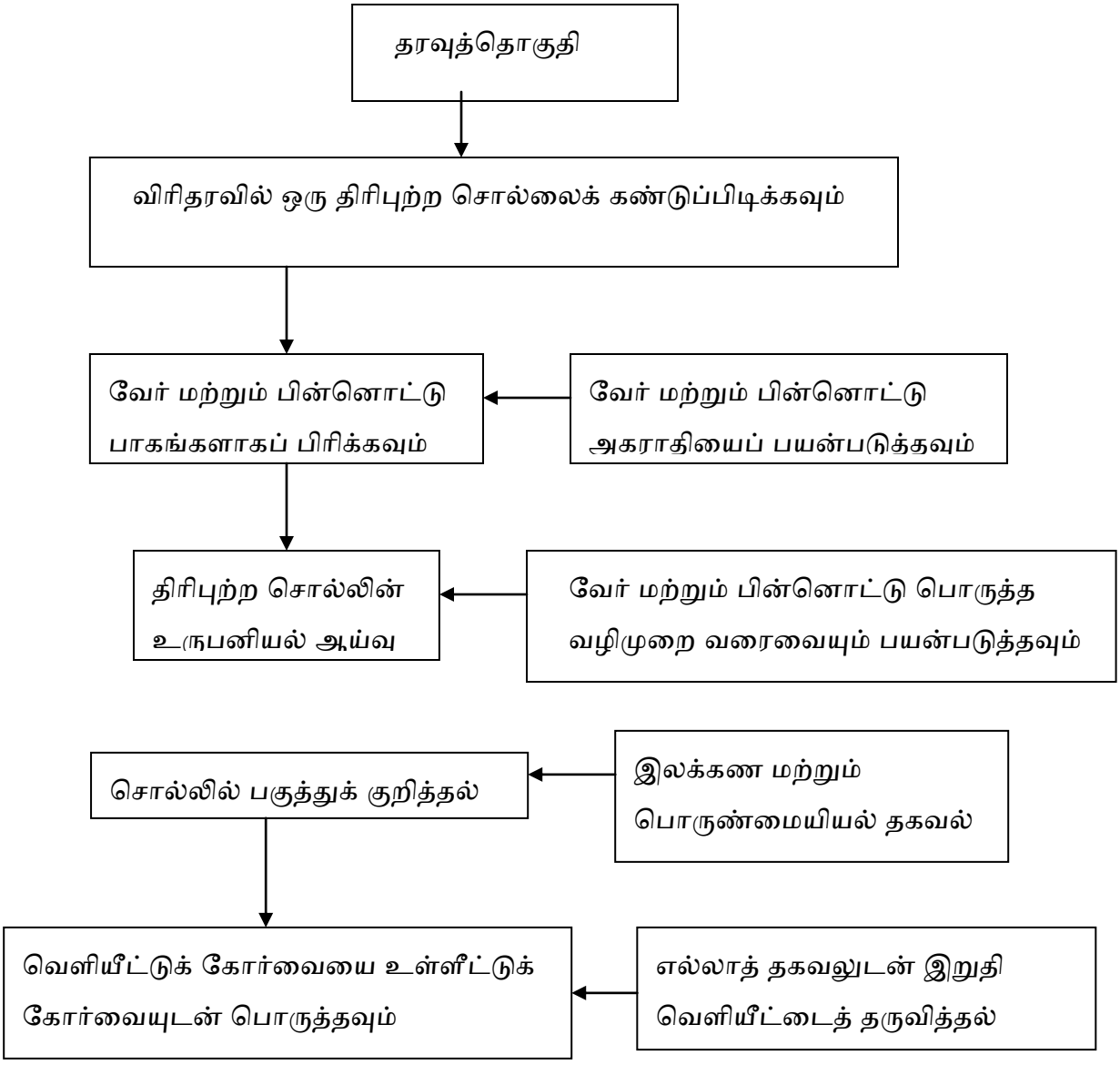
4.5 திரிபுறாத சொற்களின் ஆய்வு

திரிபுறாத சொற்களின் ஆய்வு (Processing Non-Inflectecd Forms) பின்வரும் படம் சொற்களின் பகுப்பாய்வை விளக்கும்.


தரவுத்தொகுதி
சொல்லின் கண்டுபிடித்து விதியைப் பயன்படுத்தி உரையில் ஒரு சொல்லைத் தேடவும்
எழுத்துக்களின் ஒரு கோர்வையை ஒரு சொல்லாகக் கண்டுக்கொள்ளவும்
அச்சொல்லை வேர் அகராதியில் தேடவும்
கண்டுப்பிடிக்கப்பட்டால், பொருண்மைக்கும் பிற தகவல்களுக்கும் வேண்டி இலக்கண அகராதியில் தேடவும்
கண்டுபிடிக்கப்படாவிட்டால் அதைத் திடிபுற்ற சொல்லாகக் கருதவும் அதபடி முன்னேறவும்
அடுத்த சொல்லி நேரிடவும்
அடுத்தச் சொல்லை நேரிடவும்
விளைவைத் தரும்

4.6 திரிபுற்ற சொற்களைப் பகுத்தாய்தல்

பின்வரும்படம் திரிபுற்ற சொற்களின் (Processing Inflecled words) பகுப்பாய்வை விளக்கும்.


தரவுத்தொகுதி
விரிதரவில் ஒரு திரிபுற்ற சொல்லைக் கண்டுப்பிடிக்கவும்
வேர் மற்றும் பின்னொட்டு பாகங்களாகப் பிரிக்கவும்
வேர் மற்றும் பின்னொட்டு அகராதியைப் பயன்படுத்தவும்
திரிபுற்ற சொல்லின் உருபனியல் ஆய்வு
வேர் மற்றும் பின்னொட்டு பொருத்த வழிமுறை வரைவையும் பயன்படுத்தவும்
சொல்லில் பகுத்துக் குறித்தல்
இலக்கண மற்றும் பொருண்மையியல் தகவல்
வெளியீட்டுக் கோர்வையை உள்ளீட்டுக் கோர்வையுடன் பொருத்தவும்
எல்லாத் தகவலுடன் இறுதி வெளியீட்டைத் தருவித்தல்

4.7 இரட்டை சொற்களைப் பகுத்தாய்தல்

இரட்டைச் சொற்களின் ஆய்வு (processing dbouble words) உறுப்புகள் அவற்றிற்கு இடையில் ஒரு இடைவெளிப்யால் பிரிக்கப்பட்ட கூட்டுகள் இரட்டுற்ற சொற்கள், வேறுபட்ட சொற்கள் என்பனவற்றை உள்ளடக்கும். இரட்டைசொற்களை விரிதரவில் ஆய்வது மிகவும் சிக்கலான செயல்முறையாகும். பெரும்பாலும்

இரட்டைசொற்கள் அகராதியில் தரப்பட்டிருக்கும். அவ்வாறு தரப்படாத சொற்களை காண்பது தான் சிரமமாகும். இதற்கு சேர்ந்து வருகை நிகழ்வெண்கள் இதற்கு பொருந்தும். அடுத்து வருகின்ற பொருண்மை இரட்டைச் சொற்களின் பொருண்மையை உறுதிசெய்ய உயதவுவதால் இரட்டை சொற்களை ஆய்தல் சொல் மட்டத்தில் பொருண்மை மயக்கம் தீர்க்க உதவும்.

5 தரவுத்தொகுதி அடையாளப் படுத்துகை

குறிப்பிட்ட பண்புக் குறிப்பிட்டவேண்டிய சொற்களுக்குத் தனிப்பட்ட குறியங்களை ஒட்டுவதை உள்ளடக்கும். மொழிக் குறிப்புரையின் சில வகைகள் குறிப்புரை, குறியங்கள் என்று அறியப் படுவதற்குப் பதிலாக டஇஇபரளைவஇஉ யஇஇடிவயவஇடிஇ அடையாளப்படுத்துதல் என்று அறியப்படுகிறது.

5.1 சொல் அடையாளப்படுத்துதல்

சொல் வகைப்பாடு அடையாளப்படுத்தும் (Part Of Speech Tagging) திட்டம் ஒரு சொல்லின் வாக்கியத்தில் அதன் சொல் வகைப்பாட்டுடன் அடையாளப்படுத்தும். இது போன்ற நிலைகளில் செய்யப்படுகிறது. முன்திருத்தல், தானியங்கு அடையாளம் தருகை மற்றும் மனித முயற்சிசார் பின்திருத்தல் என்பவையாகும்.

5.2 இலக்கண அடையாளப்படுத்தல்

சொல் வகைப்பாடு அடையாளப்படுத்தலில் இரண்டாவது மட்டம் இலக்கண அடையாளப்படுத்தல் ஆகும். இது தரவுத்தொகுதியில் உள்ள ஒவ்வொரு சொல்லுக்கும் இலக்கணம் பொருள் தரும் விரிவான குறிப்புரை ஒழுங்கு முறையாகும். இம்முறையில் ஆங்கிலத்திற்கு சில கருவிகள் உருவாக்கப்பட்டுள்ளது. எம்முறையில் செய்திருதாலும்

இறுதியில் மனித முயற்சிசார் திருத்தலுக்கும் மதிப்பீட்டிற்கும் திட்டவரைவு செய்யப்பட்டுள்ளது.

5.3 சொல் பொருள் அடையாளப் படுத்தல்

சொல் பொருள் அடையாளப்படுத்தல் (word sense tagging) சொல் அடையாளப்படுத்தப்பட்ட மற்றும் இலக்கணம் அடையாளப்படுத்தப்பட்ட உள்ளீட்டு உரையின் சொற்களை ஏற்க வல்லது. தானியங்கு அடையாப் படுத்தலுக்குப்பின் ஒவ்வொரு சொல்லும் சரியான பொருண்மை வகுப்பாக்கத்தை உறுதி செய்யவேண்டி மனித முயற்சி சார் பின் திருத்தல் செய்யப்படுகிறது.

6 மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதி

மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பு இன்றைய காலகட்டத்தில் மிக சிறப்பாகவும் பரவலாகவும் பேசப்படும் மற்றும் பயன்படுத்தப்படும் நெறிமுறையாகும். மொழி புரிந்துகொள்ளுதல், பேச்சு புரிந்துக்கொள்ளுதல், உரை மீட்பு மற்றும் புரிந்துக்கொள்ளுதல் உரைகளிலிருந்து தகவல்களை மீட்டல், ஒலிவழி எழுத்து புரிந்துக்கொள்ளுதல் இயந்திரமொழிபெயர்ப்பு போன்றவற்றை உள்ளடக்கும். கணிப்பொறி அடிப்படையிலான மொழியியல் தொழில்நுட்பத்தில் ஆர்வம் கூடி வருகிறது. இருப்பினும் மனித மொழியியல் உட்படும் கலவைத்தன்மைகள் மற்றும் சிக்கல்கள் காரணமாக மொழி ஆய்வுக்காகத் திட்டமிடப்படும் கணினி வழியமைப்பு முறைகள் வேறுபட்ட மொழித் தரவுகளின் மிகக் கூடுதலான அளவை வேண்டும். இதன் காரணமாக அளவீட்டு மொழியியல் என்று அழைக்கப்படும் தரவுத்தொகுதி மொழியியல் பயன்பாட்டு மொழியியலின் புதிய கிளையாக உருவாக்கியுள்ளது. எல்லா மொழிகளிலும் தரவுத்தொகுதி உருவாக்கப்பட்டு அவற்றை மொழியின் பல நிலைகளிலும் அடையாளப்படுத்தும் முயற்சி நடைப்பெற்று வருகிறது. இத்தகைய தரவுத்தொகுதி

மொழிக் கருவிகள் தயாரிப்பதற்கு மிகப்பயனுள்ளதாக அமைகிறது. இந்நெறிமுறை மொழித் தொழில் நுட்பமாக வளர்ந்துள்ளது.

6.1 மொழித்தொழில் நுட்பத்தில் தரவுத்தொகுதியின் முக்கியத்துவம்

வேறுபட்ட வகையானன தரவுத்தொகுதி உருவாக்கத்தால் மொழித்தொழில் நுட்ப ஆய்வுப் பயன்பாட்டிலும் தரவுத்தொகுதியின் முக்கியத்துவம் பெருகியுள்ளது. மொழி ஆய்வுக்கான கருவிகள் மற்றும் ஒழுங்குமுறைகள் உருவாக்குவதில் மொழி தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி பெரிதும் பங்களிப்பு செய்கிறது. பயன்பாட்டு அடிப்படையிலான தொழில் நுட்பச் செயலுக்கு தரவுத்தொகுதி உயர்ந்த விளைவைத் தருகிறது. பொதுவாக நாம் தரவுத்தொகுதியின் பயன்பாட்டை இயந்திரத்தால் கட்டுப்படுத்தப்பட்ட கருவிகளைத் திட்டமிடுவதற்கும் தானியங்கு கருவிகளைப் பரிசோதிப்பதற்குப் பயிற்சி தரவும் சிறப்பான மூலவளமாக காண்கிறோம். விரிதரவிலிருந்து பெறப்பட்ட மொழிப் பண்புக்கூறுகளின் நிகழ்வெண்ணிக்கை மொழி கற்போருக்கு நூல்களைத் திட்டமிடவும் ஒளப்வழி எழுத்துக்களைப் புரிந்துகொள்ளும் (OCR) ஒழுங்குமுறைகளை உருவாக்குவதற்கும் தானியங்கு எழுத்துப்பிழைத் திருத்தங்களை உருவாக்குவதற்கும் பயனுள்ளதாக இருக்கிறது. குறிப்புரை செய்யப்பட்ட மற்றும் குறிப்புரை செய்யப்படாத தரவுத்தொகுதிகள் இயந்திரமொழிப்பெயர்ப்பு ஒழுங்குமுறையை திட்டமிடவும் இயந்திரத்தால் படிக்கப்பெறும் அகராதிகளைத் (Machine Readable Dictionaries (MRDS)) உருபனியல் ஆய்விகள், சொல் ஆய்விகள், ஆய்விகள், வாக்கியப் பகுத்துக்குறிப்பான்கள் என்பனவற்றை உருவாக்கவும் பயனுள்ளதாக அமைகிறது. கூப்ரிதரவைப் பயன்படுத்தி உருவாக்கப்பட்ட பல கணினி கருவிகளும் ஒழுங்குமுறைகளும் மொழியைப் பயன்படுத்துபவர்களும், ஆய்வாளர்களுக்கும், எழுத்தாளர்களுக்கும, கல்வியாளர்களுக்கும், ஆசிரியர்களுக்கும், மாணவர்களுக்கும், அறிஞர்களுக்கும், அச்சிட்டு வெளிப்யிடுபவர்களுக்கும், மொழி கற்பவர்களுக்கும் மற்றும் பிறருக்கும் மிகுந்த பயனுள்ளதாக அமைகிறது. பொதுவாக தரவுத்தொகுதியின்

முக்கியதுவத்தை மொழித் தொழில்நுட்பத்தின் நான்கு விரிந்த களங்களாகப் பகுக்கலாம் : இது அவற்றின் இயல்பு மற்றும் பயன்பாட்டு நோக்கு அடிப்படைடயில் செய்யப்படும்.

- அறிவின் மூலவளமாக தரவுத்தொகுதி
- மொழித் தொழல்நுட்ப கருவிகளையும் ஒழுங்குமுறைகளையும் திட்டமிட மூலவளமாக தரவுத்தொகுதி
- மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளின் மூலவளமாக தரவுத்தொகுதி
- மனித இயந்திர இடைமுக (இடையாக) ஒழுங்குமுறைகளின் மூலவளமாக தரவுத்தொகுதி
- அறிவின் மூலவளமாக தரவுத்தொகுதி
- அறிவு மூலவளமாக தரவுத்தொகுதி பின்வரும் செயல்பாடுகளுக்குப் பயன்படுகிறது
- பன்மொழி நூலகங்களை உருவாக்குதல்
- மொழி கற்பவர்களுக்குப் பாடநூல்களைத் திட்டமிடல்
- (அச்சிட்ட மற்றும் மின் வடிவ) ஒரு மொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) இருமொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) பன்மொழி அகராதிகளை உருவாக்குதல்
- (அச்சிட்ட மற்றும் மின் வடிவ) ஒருமொழி சொற்களஞ்சியங்களைப் (பொருட்புல அகராதிகளை) உருவாக்குதல்
- பல்வேறு விதமான (அச்சிட்ட மற்றும் மின் வடிவ) நோக்கீட்டுப் பொருள்கள் உருவாக்குதல்
- இயந்திரத்தால் படிக்கவியலும் அகராதிகளை உருவாக்குதல்
- பன்மொழி சொல் மூலவளங்களை உருவாக்குதல்
- மின் அகராதிகளை உருவாக்குதல்

6.2 மொழித் தொழில் நுட்பக் கருவிகளைத் திட்டமிடுதல்

மொழித் தொழில் நுட்பக்கருவிகள் பின்வருவனவற்றை உள்ளடக்கும் :

- சொல்லாய்வு ஒழுங்குமுறை
- எழுத்துப்பிழைத்திருத்தம் ஒழுங்குமுறை
- உரை நேர்செய்யும் ஒழுங்குமுறை
- உருபனியல் பகுப்பாய்வுஒழுங்குமுறை
- வாக்கியப் பகுத்துக்குறிப்புஒழுங்குமுறை
- நிகழ்வெண் கணக்கிடும் ஒழுங்குமுறை
- சொல் தேடும் இயந்திரம்
- உரை சுருக்கும் ஒழுங்குமுறை
- உரை அடையாளப்படுத்தும் ஒழுங்குமுறை
- தகவல் மீட்கும் ஒழுங்குமுறை
- தொடரடைவு ழுங்குமுறை
- சொற்பொருள் மயக்கம் நீக்கும் ஒழுங்குமுறை
- சொல் வலை திட்டமிடல்
- பொருண்மை வலை திட்டமிடல்
- சொல் வகைப்பாடு அடையாளப்படுத்தும் ஒழுங்குமுறை
- ஒரிடம்சார் சொற்களைக் குழுமும் ஒழுங்குமுறை

7 மொழிபெயர்ப்புக்கு உதவும் ஒழுங்குமுறைகளுக்கு மூலவளமாக தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும் :

1.மொழி மூலவளம்பெறும் ஒழுங்குமுறை

2.இயந்திர மொரிபெயர்ப்புஒழுங்குமுறை

3.பன்மொழிபெயர்ப்புஒழுங்குமுறை

4.மொழி கடந்த தகவல் மீட்வுஒழுங்குமுறை

3.8 மனித-இயந்திர இடைமுக ஒழுங்குமுறைகளுக்கு மூலவளமாக தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும்:

1. ஒலியால் வருடி எழுத்துக்களைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்
2. குரலைப் புரிந்துக்கொள்ளும் ஒழுங்குமுறைகள்
3.உரையிலிருந்து பேச்சு ஒழுங்குமுறைகள்
4.இணையவலை (இணையத்தளம்) அடிப்படையிலான கற்றல் ஒழுங்குமுறைகள்
5. கேள்விப் - பதில் ஒழுங்குமுறைகள்
6. கணினியின் உதவியுடன் கட்டளைகள்
7 .கணினியின் தவியுடன் மொழிக் கல்வி
8. உரை உருவாக்கம்

9 பேச்சுத் தொழில்நுட்பத்தில் தரவுத்தொகுதி

இது பின்வருவனவற்றை உள்ளடக்கும் :

1. பேச்சுத் தொழில் நுட்பத்திற்குப் பொதுவான சட்டகத்தின் உருவாக்கம்
2. கிளைமொழிகளில் ஒலியியல், சொல்லியியல் மற்றும் உச்சரிப்பியல் வேறுபாடுகள்
3. தானியங்கு பேச்சு புரிந்துகொள்ளுதல் செயல்பாடு
4 .தானியங்கு பேச்சு உருவாக்கும் ஒழுங்குமுறைகள்
5. தானியங்கு பகுப்பாய்வு ஒழுங்குமுறைகள்
6. தானியங்கு பேசுபவரை அடையானம் காணும் ஒழுங்குமுறைகள்
7. பேச்சுக் குறைபாடுகள் சரிசெய்தல்

10 இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி

இயந்திர மொழிபெயர்ப்பு செயற்கை அறிவு நுட்பத்தின் சிக்கலான ஒரு பகுதியாகும். இது மூல மொழியின் வாக்கியங்களிலிருந்து, இலக்குமொழி வாக்கியங்களை உருவாக்குவதற்கான ஒழுங்குமுறையை உருவாக்கவேண்டும். இலக்கு மொழியின் வெளிப்யீடு உயர்ந்த தரமுடையதாக அமையத் தேவையில்லை. ஓரளவுக்குத் தரம் வாய்ந்ததாய் இருந்தால் போதும். இயந்திர மொழிபெயர்ப்பு மொழிகளுக்கிடையே கருத்துப்பரிமாற்றத்திறகும் தகவல் பரிமாற்றத்திற்கும் மொழிகளைக் கடந்த தகவல் மீட்புக்கும் மொழி கற்றலுக்கும் முக்கியத் தொழில் நுட்பமாகும். பல விதமான பயன்பாட்டின் சாத்தியம் சாதாரணமாக இயந்திர மொழிபெயர்ப்பு மின் வணிகம் தகவல் வட்டாரமாக்கம் பல்மொழி ஆவணப்படுத்தல் மற்றும் தகவல் தொழில் நுட்பம் போன்றவற்றில் மிகப் பயனுள்ள தொழில் நுட்பமாக விளங்குகிறது. நாம் இயந்திர மொழிபெயர்ப்புத் தொழில் நுட்பத்தில் தரவுத்தொகுதியின் பங்களிப்பைப் பற்றிக் கூறுவதற்கு முன்பின்வருவரும் கேள்விகக்குத் திருப்பதிகரமான விடைதர வேண்டும். மனிதர்கள் மொழி பெயர்ப்பைச் செய்யும்போது ஏன் இயந்திர மொழி பெயர்ப்பை நாம் ஏன் உருவாக்க வேண்டும். இக்கேள்விக்கான பதில் எளிமையானது அல்ல. ஹட்சின் (Hutchin's 1986) பின்வரும் காரணங்களைத் தருகிறார்:

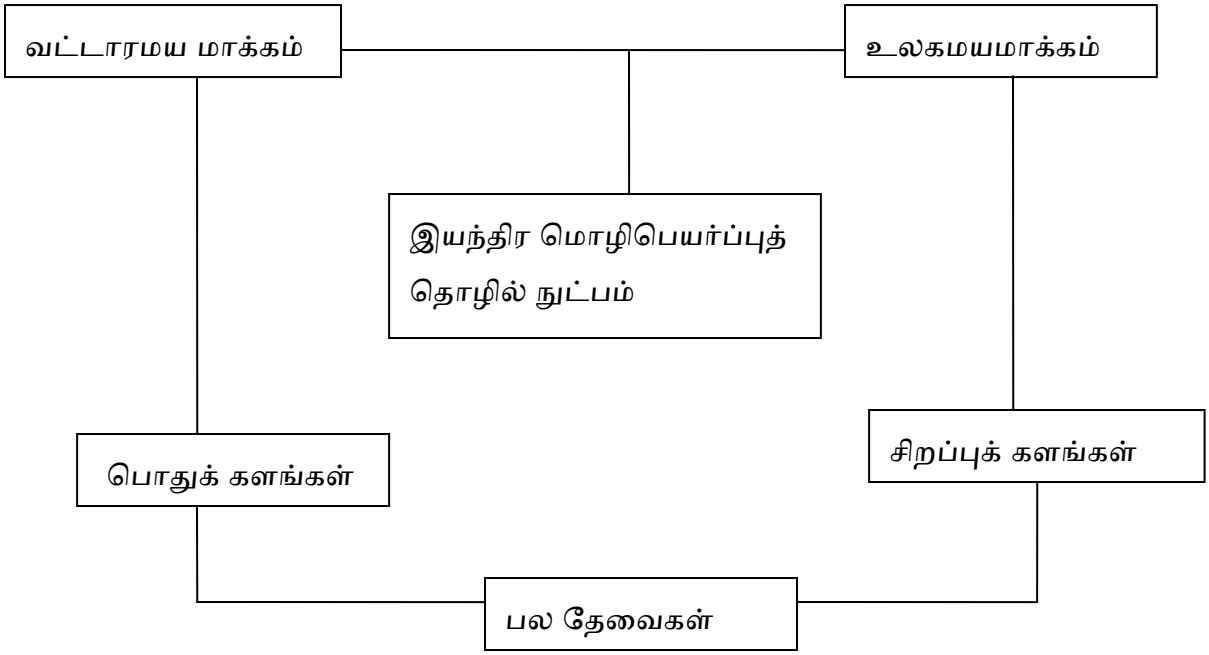
1. பல்வேறுபட்ட அறிவியல் தொழில் நுட்ப ஆவணங்கள் உலகிலுள்ள தொழில் நுட்பவியலாருக்கும் அறிவியலாருக்கும் உடனடியாகத் தரவேண்டுமானால் இயந்திர மொழிபெயர்ப்பு மிகவும் அவசியமாகும்.
2. மனித மொழி பெயர்ப்பாளர்கள் இல்லாத சூழல்களில் இயந்திர மொழிபெயர்ப்பு மிகவும் பயனுள்ளதாக அமையும்.
3. இயந்திர மொழிபெயர்ப்பு மொழகளைக் கடந்த ஆவண மாற்றங்கள் மூலம் அக நிலையான கூட்டுறவை மேம்படுத்த இயலும்.
4. இயந்திர மொழிபெயர்ப்பு பல்வேறுபட்ட அறிவியல், தொழில்நுட்பம், விவசாயம் மற்றும் மருத்துவ தகவல்களை ஏழ்மையான மற்றும் முன்னேறுகின்ற

நாடுகளுக்கு விரைவான, எளிதான மற்றம் மலிவான பரிமாற்றம் அடிப்படையில் மொழித் தடைகளை நீக்கி அக நிலையான கூட்டுறவையும் மேம்படுத்த இயலும்.

5. மொழிப் பெயர்ப்பு இராணுவ நோக்கங்களுக்கு தொழில்நுட்ப ஆய்வு மற்றும் வணிக நோக்கங்களுக்கு மிகவும் பயனுள்ளது.

1980-களில் இயந்திர ஹட்சின் இயந்திர மொழிப்பெயர்ப்புக்குச் சாதகமாகக் கூறியபோது இருந்த சூழல் கடந்த ஆண்டுகளில் மிகவும் மாறியுள்ளது. இந்நூற்றாண்டின் தொடக்கத்தில் மக்களின் வாழ்க்கை நடத்தை காரணமாக தாய்மொழி மூலம் தகவலைப் பெறுவது விதியாக மாறிறயுள்ளது. மேலும் படிப்பறிவின் வளர்ச்சி, செய்தி மற்றும் தகவலின் உலகமயமாக்கம், நாடுகளைக் கடந்த பன்மொழிய தரவுத்தொகுதி, இனங்களுக்கிடையே மொழித் தனித்தன்மையின் அவசியம் மற்றும் பிற காரணிகள் கணினி அறிவியலாரையும் மொழியலாரையும் இயந்திர மொழிப்பெயர்ப்பு ஒழுங்குமுறையை உரவாக்க இணைந்து பணியாற்றும் சூழலுக்குத் தள்ளியுள்ளது.

இக்காரணிகள் பங்களிப்பைப் பின்வரும் வரைபடத்தின் மூலம் சுருக்கமாக கூறலாம்.



இங்கு அனுபவவாத தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு முன்மொழியப்படுகின்றது. இது மொழிபெயர்ப்பில் சூடுபடுத்தபட்டடுள்ள மொழிகளின் உரைகளிலிருந்து உருவாக்கப்பட்ட மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதிகளின் பகுப்பாய்விலிருந்து பெறப்பட்டத் தகவல்களின் மற்றும் எடுத்துக்காட்டுகளின் அடிப்படையில் அமைந்தது. இங்கு இயந்திர உதவியால் செய்யப்படும் மனித மொழிபெயர்ப்பிற்கு மொழிபெயர்ப்பு தரவுத்தொகுதி பயன்படுத்தப்பட்ட விவாதிக்கப்படுகின்றது. எல்லைப்படுத்தப்பட்ட மற்றும் பொதுக் களங்களில் இவ்வொழுங்குமுறையின் பயன்பாடு அழுந்தக்கூறப்படுகிறன்றது. தகவல் பரிமாற்றத்தின் உலக உயர்வுக்கும் வட்டாரவாக்கத்திற்கும் மொழிபெயர்ப்பு ஒழுங்குமுறையின் தேவை விளக்கப்பட்டு எதிர்கால முன்னேற்றத்திற்கு வேண்டிய வழிமுறைகள் முன்வைக்கப்படுகின்றன. மொழி பெயர்ப்பு வரலாற்றின் மற்றும் தோல்வியல் இன்று கற்றுக்கொண்டவைகளின் அடிப்படையில் இயந்திர மொழிபெயர்ப்பு அமைக்கப்பட்ட ஆலோசனை கூறப்பட்டுள்ளது.

10.1 நோக்கம்

கடந்த 50 ஆண்டுகளில் இயந்திர மொழிபெயர்ப்பு தொழில் நுட்பங்கள் மற்றும் பயன்பாடு இவற்றின் அடிப்படையில் கூறத்தக்க அளவு வளர்ந்துள்ளது.இன்றைய இணைய தள கால கட்டத்தில் இணைய தகளத்தில் தகவல்கள் பெரியளவில் பல மொழிகளில் உள்ளன. அவற்றை எல்லா மொழி பேசுவோரும் பகிர்ந்து கொள்ளவேண்டுமானால் அவை ஒரு மொழியிலிருந்து மற்றொரு மொழிக்கு உடனடியாக மொழி மாற்றம் செய்யப்படவேண்டும். இதை இயந்திர மொழிபெயர்ப்பின் விரைவான முன்னேற்றத்திற்ன் அடிப்படையில் தான் நிறைவேற்ற இயலும். இது நடைமுறையில் சாத்தியமில்லாத செயல்பாடாக தற்போது தோன்றினாலும் நாம் இதைச் சாத்தியமாகச் செய்கின்ற இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை உருவாக்குவதில் ஆவர்ம் காட்டவேண்டும்.

இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறையை ஒரு வடிவிலிருந்து மற்றொருவடிவுக்கு (அதாவது சொல்லிலிருந்து சொல்லையோ கூட்டுச் சொற்களிலிருந்து சுட்டுச் சொற்களையோ தொடரிலிருந்து தொடரையோ வாக்கியத்திலிருந்து வாக்கியத்தையோ) இடப்பெயர்ச்சி செய்யும் அல்லது ஒரு மொழியிலுள்ள உரைகளை மற்றொரு மொழியில் உருப்படுத்தம் செய்யும் எளியமுறை அல்ல. இயந்திர மொழிப்பெயர்ப்பு உருவாக்கப்பட்ட உரை இலக்குமொழியில் இலக்கண அடிப்படையில் சரியான மற்றும் கருத்துரு அடிப்படையில் ஏற்றுக் கொள்ளத்தக்க வகையில் அமைய உறுதி தருவதைக் குறிப்பிடுகின்றது. இயந்தி மொழிப்பெயர்ப்பு அர்த்தம் மற்றும் பொருளடக்க அடிப்படையில் ஒன்றாக இல்லாவிடினும் குறைந்தது மூலமொழிக்கு அண்மையில் இருக்கவேண்டும். மூல மொழியிலுள்ள தகவல் இலக்கு மொழியில் மாற்றப்படும் போது இழப்பு ஏற்படக்கூடாது. மொழிப்பில் மூலமொழியை விடக்கூடுதல் தகவல் சேர்க்கப்படலாகாது. இயந்திர மொழிபெயர்ப்பு உருபனியல், சொல்லியல், தொடரியல் பொருண்மையியல், பயன்வழியியல், கருத்தாடல், புலனறிவு இவற்றைச் சார்ந்திருக்கவேண்டும். இதன் அர்த்தம் என்னவென்றால் ஒரு இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை ஒரு மனித மொழிபெயர்ப்பாளருக்குள்ள திறனைப் பெற்றிருக்கவேண்டும் என்பதாகும். இதைச் சாதிப்பதற்கு நாம் காத்திருக்கவேண்டும். இருப்பினும் இருமொழி

அல்லது இணை மொழி பெயர்ப்பு தரவுத்தொகுதிகள் இதைச் செய்வதற்குச் சாத்தியமான சூழலைத் தருகிறது. இது நம்மை தரவுத்தொகுதிகள் அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு அணுகுமுறைக்குக் கொண்டு செல்கிறது. இதன் காரணமாக இயந்திர மொரிப்பெயர்ப்பு ஒழுங்குமுறையின் திறன் கூடுகிறது. குறையான தகவலிலன் மூலமாகத் திருத்தப்படாத வடிவில் பயன்படுத்தப்படுவதன் காரணமாகவும் இயந்திர மொழிப்பெயர்ப்பு முற்றிலும் சரியான மொழிப்பெயர்ப்பை உருவாக்குவதை நோக்கமாகக் கொண்டதல்ல எனக் கூறவியலும். சூயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறை மொழியில் அடிப்படையிலும் புலனறிவு அடிப்படையிலும் உத்தேச ஒழுங்குமுறையைத்தான் நோக்கமாகக் கொள்கிறது. முயற்சித்தல் மற்றும் பிழை செய்தல் என்ற நீண்ட செயற்பாங்குளின் வழி இலக்கை அடையலாம்.

10.2 வரலாற்றிலிருந்து கிடைக்கின்ற பாடம்

1950-களில் இயந்திர மொழி பெயர்ப்பு உலகிலுள்ள பெருமாபாலான எல்லா மொழிகளுக்கும் நேரடியாகவோ மறைமுகமாகவோ பல மூல மொரியிலிருந்து இலக்கு மொழிக்கு இயந்திர மொழிபெயர்ப்பு செய்ய முயற்சிகள் எடுக்கப்பட்டன. இது வரை முழு வெற்றி அடைய இயலவில்லை. அதன் அர்தம் என்னவென்றால் மூல மொழிக்கும் இலக்கு மொழிக்கும் மொழி பயன்பாட்டாளர்களால் ஏற்றுக்கொள்ளக்கூடிய மொழிப்பெயர்புகளைத் தானியக்கமாக உருவாக்கும் ஒரு இயந்திர மொழிபெயர்ப்புப் ஒழுங்குமுறை இல்லை. இதன் விளைவாக இயந்திர மொரீபெயர்ப்பு ஆய்வாளர்கள் மரபு சார்ந்த விதி அடிப்படையிலான அணுகுமுறைக்குத் தங்கள் கவனத்தை திருப்பியுள்ளானர்.

.10.3 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறை

கடந்த நூற்றாண்டுகளாக இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகளுக்கு வேண்டி உருவாக்கப்பட்ட விதிகளின் குழுமங்களால் இயற்கை மொழியின் நெகிழ்வான அமைப்பை உருப்படுத்தம் செய்ய இயலாது. இதன் காரணமாக தரவுத்தொகுதி

அடிப்படையிலான இயந்திர மொழி பெயர்ப்பு உருவானது. இது பின்வரும் இயந்திர மொழிபெயர்ப்பு வகைகளை உள்ளடக்கும்.

1. எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (example based machine translation)
2. புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு (statics based machine translation)

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் விளைவுகளில் ஒன்று இருமொழி உரைகளின் வாக்கியங்களை வரிசைப்படுத்தும் வழிமுறை வரைவுகளின் உருவாக்கமாகும். இவ்வெளிப்ய விளைவு இயந்திர உதவி வழி மனித மொழி பெயர்ப்பின் அடிப்படையில் விக்ஷயங்களில் ஒன்றாக அமைந்தது. ஏனென்றால் இது மொழிபெயர்ப்புக்கு உதவும் பல புதிய கருவிகளுகடகுப் பொருத்தமான அடிப்படையாக அமைந்துள்ளது. மொழிபெயர்ப்பாளர்களால் ஏற்கனவே உருவாக்கப்பட்ட மொழிபெயர்ப்புகளைப் பயன்படுத்தி முழுமையாகவோ பகுதிகாகவோ அவற்றின் அக அமைப்புகளைக் கண்டுபிடிக்க முயலுகிறது. இப்பகுப்பாவு அடிப்படையிலான அணுகுமுறை மொழி பெயர்ப்பாளர்களுக்கு உதவும் கருவிகளைஉருவாக்க உதவியது. இருமொழி இல்லது இணைமொரி பெயர்ப்பு விரிதரவை இயந்திர மொழிபெயர்ப்பில் பயன்படுத்தும் கருத்து புதியது அல்ல. இயந்திர மொழிபெயர்ப்பின் ஆரம்ப காலத்திலேயே இது முயற்சிக்கப்பட்டது. பொதுவாக இணை தரவுத்தொகுதிகள் ஒரு மொழி தரவுத்தொகுதிகளை விட மொழிகளைப் பற்றிய விரிவான தகவலைக் கொண்டிருக்கிறது. தரவுத்தொகுதி ஒழுங்குமுறை எடுத்துக்காட்டு அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு அணுகுமுறையையும் புள்ளியியல் அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அணுகுமுறையும் இணைத்து பயன்படுத்துவதால் இது உயர்ந்ததாக இருக்கிறது. விரிததரவு அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு மூல மொழியிலும் இலக்கு மொழியிலும் உள்ள தரவுத்தொகுதிகளில் முதன்மையான பகுப்பாயிவிலிருந்து உருவாக்கப்பட்ட வகைகளின்

பரப்பெல்லைகளின் அடித்தளத்தில் அமைந்துள்ளது. பகுப்பாய்வு மொழிபெயர்க்கப்பட்ட விரிதரவில் அடங்கியுள்ள தொடர்கள், மரபுத் தொடர்கள், வாக்கியங்கள், பத்திகள் மற்றும் பிற மொழியியல் பண்புக்கூறுகளின் உருபனியல் பொருண்மையியல் மற்றும் புலனிறவு பொருள்கோள்களை உட்படுத்தும்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் அடிப்படையிலான நெறிமுறை மொழி பெயர்ப்புக்கு முன் வரையறுக்கப்பட்ட தீர்மானங்கள் இல்லை. ஆனால் மனிதரால் மொழி பெயர்க்கப்பட்ட உரைகளில் மிக சாத்தியமான தீர்மானங்கள் காணப்படும் என்ற நம்பிக்கையின் அடிப்படையில் அமைவதாகும். அதாவது மனித மொழிபெயர்ப்பாளர்கள் திறமையின் பெரும்பகுதி மொழி பெயர்க்கப்பட்ட உரைகளில் காணப்படும் மொழி நிகரன்களில் குறியாக்கம் செய்யப்பட்டுள்ளன என்று கூறலாம். எல்லைக்குட்பட்ட களங்களில் இவ்வணுகுமுறையால் பெறப்பட்ட சமீபகால வெற்றி பொதுவான களங்களிலும் இம்மாதிரியான வெற்றயைப் பெற மொழி சார்ந்த மற்றும் மொழிக்குப் புறம்பான அறிவுதேவை என்று காட்டுகினறது. எல்லாக் களங்களிலும் இவ்வணுகுமுறையின் வெற்றி குறித்து தீர்மானமாக வருவதுரைக்க இயலாது. இருப்பினும் நன்றாகத் திட்டமிடப்பட்டது மொழிப்பெயர்ப்பு விரிதரவிலிருந்து பெறப்பட்ட தகவலிலிருந்து உருவாக்கப்பட்ட நெறிமுறை எல்லைக்குட்பட்ட மற்றும் பொது வளங்களில் வெற்றயைத் தரும் என்று கூறு இயலும்.

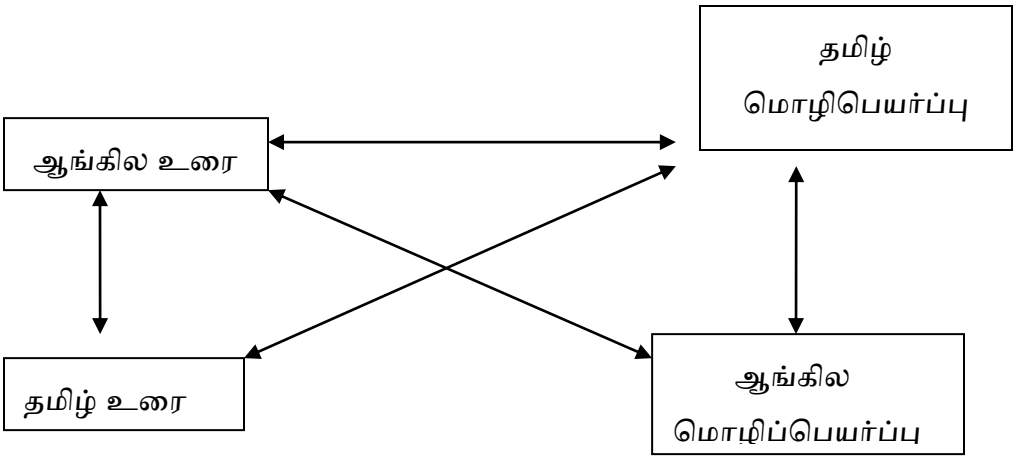
10.4 தரவுத்தொகுதி அடிப்படையிலான அணுகுமுறையுடன் தொடர்புடைய சிக்கல்கள்

தரவுத்தொகுதி அடிப்படையிலான மொழிபெயர்ப்பு விரிவான தரவுத்தொகுதி பகுப்பாய்வின் மற்றும் ஆய்வின் கூறத்தக்க முன்னேற்றத்தைப் பெற்றுள்ளது. இந்நாள் வரை பல்வேறு வகையிலான மொழிபெயர்ப்பு ஒழுங்குமுறைகள் திட்டமிட உதவும் முக்கியமான உள்ளறிவைத் தரவேண்டி உருவாக்கப்பட்டுள்ளன.. பொதுவாக இம்மொழிபெயர்ப்பு தரவுத்தொகுதி இயற்கையாகப் பெறப்படும் மொழித் தரவுகளின்

பெரிய சேகரிப்பை உருப்படுத்தும் செய்கின்றது. இயந்திர மொழிபெயர்ப்பு ஒழுங்குமுறைகள் மொழிபெயர்ப்பு தரவுத்தொகுதியின் பயன்பாடு கட்டாயமானதாகும் என்று தெளப்வாக்குகின்றது. ஏனென்றால் அவை மொழி மற்றும் மொழிக்குப் புறம்பான எடுத்துக்காட்டுகளின் மற்றும் தகவல்களின் பல்வேறு விதமான வகைகளை நமக்குத் தருகிறது.

10.5 மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கம்

மொழிப்பெயர்ப்பு தரவுத்தொகுதி மூல மொழியின் உண்மையான பண்புக் கூறுகளையும் இலக்கு மொழியிலிருந்து பெறப்பட்ட அவற்றின் மொழி பெயர்ப்புகளையும் கொண்டிருக்கும். இவ்தரவுத்தொகுதி மொழிகளைக் கடந்து சொற்களின் மற்றும் தொடர்களின் பொருண்மையையும் செயல்பாட்டையும் கொண்டிருக்கும். இதன் காரணமாக ஒரே கட்டுபாட்டிற்குள் இரு வேறுபட்ட மொழிகள் குறிப்பிட்ட பொருண்மைகளின் மெய்யுருவாக்கத்தை ஒப்பிடுவதற்குப் பொருத்தமான அடிப்படையைத் தருகின்றது. மேலும் அவை மொழிகளைக் கடந்த வேறுபட்ட சொற்களைக் கண்டுபிடிப்பதைச் சாத்தியமாக்குகிறது. இவ்வாறு மொழிப்பெயர்ப்பு தரவுத்தொகுதி மொழிகளைக் கடந்த தரவுப் பகுப்பாய்வுக்குத் தருகின்றது. மொழிபெயர்ப்பு தரவுத்தொகுதியின் உருவாக்கம் ஒரு சிக்கலான வேலையாகும். இது தரவுத்தொகுதியின் உருவாக்கம் மற்றும் கவனமான வழிகாட்டலை வேண்டுகிறது. மொழிபெயர்ப்பு தரவுத்தொகுதி ஒப்பீட்டு இணை தரவுத்தொகுதிகளின் முன்னேற்றத்திற்கான வழிகளை ஒன்றிணைக்க வேண்டி தகுதியுடையதாக ஆக்க உருவாக்கப்படும் இரு மொழிகளிலிருந்து உரை மாதிரிகள், உரை வகை, பாடப்பொருள், நோக்கம் மற்றும் கலைச்சொல் இவற்றின் அடிப்படையில் இயன்ற வரைக்கும் பொருத்தப்படும் இரு மொழிகளுக்கிடையிலுள்ள மொழிப்பெயர்ப்பு தரவுத்தொகுதியின் அமைப்பு பின்வருமாறு அமையும்.



இம்மாதிரியான தரவுத்தொகுதிக்கு உரை மாதிரிகள் தெரிவு பின்வருமாறு கொள்கைகளால் பொதுவாக வழி நடத்தப்படும்.

1.எழுதப்பட்ட உரைகளில் பயன்படுத்தப்படும் மொழி மாதிரிகள் மட்டும் தான் மொழி பெயர்ப்பு விரிதரவில் உட்படுத்தப்பட்டுள்ளன

2.பேச்சு கருத்தாடலிலிருந்து பெறப்படும் உரைகளை உட்படுத்தும் சந்தர்ப்பம் இல்லை. ஏனென்றால் தற்போதைய இயந்திர மொழிபெயர்ப்பு எழுதப்பட்ட உரைகளை மட்டுமே இலக்காகக் கொண்டுள்ளது. உட்படுத்தப்பட்டுள்ள உரைகள் தற்கால மொழியைப் பிரதிபலிப்பது எதிர்பார்க்கப்படும். இருப்பினும், பழங்கால நூல்கள் வரலாற்று உரைகளின் மொழி பெயர்ப்புக்குத் தேவையானதாகும்

3.மொழிப்பெயர்ப்பு தரவுத்தொகுதி எந்தக் குறிப்பிட்ட வட்டார மொழிக்கோ மொழி வகைக்கோ உரிய உரை வகைக்கு எல்லைப்படுத்தப்படவில்லை. அவை மொழி பயன்பாட்டின் எல்லாச் சாத்தியமான களங்கள் மற்றும் துறைகளிலிருந்து பெறப்பட்ட உரை வகைகளின் பரந்த பரப்பெல்லையை உட்படுத்தும்.

4.இரண்டு மொழிகளிலிருந்து எடுக்கப்பட்ட உரைகள் இயன்றவரை ஒப்பிட தக்கவையாக இருக்கும். இவை இனம் (எ.கா.செய்தி), வகை (எ.கா. அரசியல்), பொருளடக்கம் (எ.கா.தேர்தல்) மற்றும் வடிவம் (எ.கா.அறிகை) இவற்றின் அடிப்படையில் பொருத்தம் பார்க்கப்படும். அவைகள் நோக்கம்

பயன்படுத்துபவரின் மொழி இவற்றின் அடிப்படையிலும் பொருத்தமுடையதாக இருக்கும்.

5.மொழிபெயர்ப்பு விரிதரவில் உட்படுத்தப்படுள்ள மாதிரிகள் ஒரு இயல்பான தடை நிலையில் (எ.கா. பகுதி, துணைப்பகுதி, அல்லது பத்தி போன்றவை) தொடக்கத்திலிருந்து இறுதிவரை எடுக்கப்பட்ட உரைகளில் விரிவான மற்றும் இயற்கையான பாகங்களைக் கொண்டிருக்கும்.

10.6 மொழிபெயர்ப்புத் தரவுத் தொகுதியைப் பொருத்தமாக வரிசைப்படுத்தல்

மொழிபெயர்ப்பு தரவுத்தொகுதிகளைப் பொருத்தமாக வரிசைப்படுத்துவது என்பது மூலமொழியின் ஒவ்வொரு மொழிபெயர்ப்பு ஆகும். இலக்கு மொழியின் ஒவ்வொரு மொழிபெயர்ப்பு அலகுடன் பொருந்தவேண்டும் என்பதாகும். இங்கு மொழிபெயர்ப்பு அலகு என்பது சிறிய தொடர்ச்சிகளான சொற்கள், கூட்டுச்சொற்கள், தொடர்கள் மற்றும் வாக்கியங்கள் என்பனவற்றுடன் பெரிய தொடர்ச்சியான பத்திகள், பகுதிகள் என்பனவற்றையும் உள்ளடக்கும். மொழிபெயர்ப்பு அலகுகளின் தெரிவு மொழியில் ஆய்வுக்குத் தெரிந்தெடுக்கப்பட்டுள்ள பார்வையையும், பயன்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகையையும் பொறுத்து அமையும். மொழி பெயர்க்கப்பட்ட தரவுத்தொகுதி மூலத்திலிருந்து (எ.கா.சட்டம் மற்றும் தொழில் நுட்ப தரவுத்தொகுதிகள்) மாறாத உயர்ந்த நிலையை வேண்டினால் இரு தரவுத்தொகுதிகளின் நெருங்கிய வரிசைப் பொருத்தம் வாக்கியங்கள், சொற்கள் என்ற அளவில் விலகி நிற்கும். தரவுத்தொகுதி தழுவலாக இருந்தால் மூலத்தின் நேரடி மொழிபெயர்ப்பாக அமையாமல் பத்திகள், பகுதிகள் என்ற பெரிய அலகுகளை வரிசைப்படுத்தப்பட்டுள்ள தரவுத்தொகுதியின் வகை அடிப்படையில் நேர்படுத்தப்படும். வரிசையாக வருவதும் மனித மொழிபெயர்ப்பாளர்களின் நம்பகத் தன்மையும் மொழிபெயர்ப்பட்ட தரவுத்தொகுதிகளை வரிசைப்பொருத்தம் செய்ய உதவும். இது கூடுதல் தொழில்நுட்ப தரவுத்தொகுதிக்குப் பகுதி உண்மையாகும். மாறாக, இலக்கிய தரவுத்தொகுதிகள், தரவுத்தொகுதியில் பயன்படுத்தப்படும் நிகரன்கள் முன்னரே முறையாக்கம் செய்யப்பட்டிருந்தால் வாக்கிய

மட்டத்திற்கு கீழே உள்ள அலகுகளின் வரிசைப் பொருத்தத்திற்கு உட்படும். எந்த நிலையிலும் (பத்தி, வாக்கியம், சொல் போன்ற நிலைகளில்) மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதிகள் இணையான அலகுகள் கொண்ட எளிய சொல் தரவு மையங்களாக் கருதப்படுகிறது. இதன் முக்கிய நோக்கம் இரண்டு மொழிகளுக்கிடையே காணப்படும் அமைப்பு ஒற்றுமைகளைக் காட்டுவதற்கு அல்ல: பயன்வழியில் அடிப்படையில் மூல உரை அலகுகளுக்கு அண்மைப்பட்டதாகக் தோன்றும் இலக்கு உரை அலகுகளைத் தேடுவதாகும். இதைக் செய்வதற்குத் தொடக்கநிலை, இருமொழிய அகராதிகளின் உதவியால் சொற்களின் ஆரம்ப வரிசைப் பொருத்தத்தைச் செய்வதாகும். இம்மாதிரியான குறை வரிசைப் பொருத்தங்கள் வாக்கிய நிலைகளில் திருப்திகரமான விளைவுகளைத் தரும்.

10.7 இந்திய மொழிகளின் இன்றைய நிலை

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பில் முக்கியமான தடைகளில் ஒன்று இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளின் உருவாக்கதபகும். ஒருமொழிய விரிதரவில் போதுமான அளவு எழுதப்பட்ட உரைகள்ள இருந்தாலும் பின்வரும் செயல்படுத்துநிலை சிக்கல்களின் காரணமாக இருமொழிய மொழிபெயர்ப்பு தரவுத்தொகுதிகளை மறுதிட்டவரைவு செய்வதற்கு ஆர்முள்ள முயற்சிகள் எடுக்கப்பட்டவில்லை.

10.8 உருவாக்கம்

இந்திய மொழிகளுக்கிடையே வேறுபட்ட வகையான. மொழிப்பெயர்ப்பு தரவுத்தொகுதி நீண்டநாள் தேவையாகும். சில ஒருமொழிய தரவுத்தொகுதிகள் ஒரேமாதிரியான திட்டவரைவுக் கொள்கைகளையும் உரை வகைகளையும் கொண்டு உருவாக்கப்பட்டுள்ளன. இவற்றை முறையாக மொழிபெயர்ப்பு தரவுத்தொகுதிகளை உருவாக்கப் பயன்படுத்தலாம்.

ஏற்றுக்கொள்ளல்: மனித முயற்சியால் உருவாக்கப்பட்ட மொழிபெயர்ப்பு தரவுத்தொகுதிகள் சில (எ.கா.) (இயவஇடிஇயட bடிடிம வசரளவ) மற்றும் ளுயாவைலய ஆஉயனநஅல போன்றவற்றால் வெளிப்யிடப்பட்ட பல உரைகள்) ஏற்கனவே உருவாக்கப்பட்டுள்ளது என்றாலும் அவை ஏற்றுக்கொள்ளக்கூடிய மின்விடிவில் இல்லை.

மாற்றம்: வேறுபட்ட மூலம், உரைவகை, மாதிரி மற்றும் வடிவமைப்பு கொண்ட பிற தரவுத்தொகுதிகள் இணை மொழிபெயர்ப்பு தரவுத்தொகுதி மாற்ற வேண்டி ஆயப்படவேண்டும்.

சுத்தமாக்கல்: மொழிபெயர்ப்பு தரவுத்தொகுதி எதிர்கால செயற்பாங்கிற்கும் ஏற்றுக்கொள்ளலுக்கு பயன்பாட்டிற்கும் வேண்டி இயந்திரப் பயன்பாட்டு வடிவமைப்பில் வைத்திருக்கமொழி வல்லுனர்களால் முறையாக தலையிடப்படவும் கட்டுப்படுத்தப்படவும் வேண்டும்.

இணைப்பொருத்தமாக வரிசைப்படுத்தல்: மொழிகளுக்குடையில் பொருத்தும் பகுதிகளை அடையாளங்கான வேண்டி மொழிபெயர்ப்பு தரவுத்தொகுதி இணைப்பொருத்தமாக வரிசைப்படத்தப்படவேண்டும்.

10.9 மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளில் மொழியின் செயல்பாடுகள்

மொழிப்பெயர்ப்பு தரவுத்தொகுதி தொகுத்தலுக்கும் வரிசைப்படுத்தலுக்கும் பின்னர் மொழியில் பகுப்பாய்வின் வேறுபட்ட நிலைக்கு உள்ளாக்கப்படுவதற்கு முன்னர் பொருத்தமுஹ்ன பல நிலைகளைக் கடக்கும். இது மொழி பெயர்ப்பு நிகரன்கள் முறையாக்கத்தின் வடிவாக்கத்தின் நிலைகளுக்கு அடிப்படையாக அமையும். பொதுவாக மொழியியல் ஆய்வு பின்வருவனவற்றை உள்ளடக்கும் :

- உருபன்களின் வடிவத்தையும் செயல்பாட்டையும் அடையாளம் காண உதவும் சொற்களின் உருபனியல் ஆய்வு
- தொடர்புள்ள தரவுத்தொகுதிகளின் வடிவத்தையும் செயல்பாடுகளையும் அடையாளம் காண உதவும் தொடரியல் ஆய்வு
- தரவுத்தொகுதிகள் பயன்படுத்தப்பட்டுள்ள சொற்களின் புறவடிவங்களுக்கு இடையில் உள்ள இடைமுகத்தை ஆயு உதவும் உருவனியில் - தொடரியல் ஆய்வு. இதன் துல்லியமான மற்றும் திறமையான ஆய்வு பகுப்பாய்வின் தன்மையையும், விரிவையும் அதிகரிக்கும்.
- சொற்கள், தொடர்கள் போன்ற அலகுகளின் பொருண்மையைக் கண்டுபிடிக்கவும் இதில் வரும் பொருள் மயக்கங்களைக் கண்டுபிடிக்கவும் உதவும் பொருண்மையில்.

10.10 மொழிபெயர்ப்பு ஆய்வு

இயந்திர மொழிப்பெயர்ப்பு ஆய்வாளருக்கு இடையிலான முக்கியமான வாக்குவாதம் மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளின் ஆய்வில் உட்படுத்தம் கலவைத் தன்மையின் நிலையைப் பற்றியாகும். பொதுவான நம்பிக்கை இயற்கையான உரைகளில் வரும் பல எண்ணிக்கையிலான மொழியியல் நடத்தைகள் ஆயப்பட்டு வெளிப்ப்படையாக உருப்படுத்தம் செய்யப்படும் வரை உயர்ந்த தன்மையுள்ள இயந்திர மொழிபெயர்ப்பு எய்த இயலாததாகும் என்பதாகும். சொல் மயக்கம் மற்றும் உறுப்பு பொருத்தம் இவற்றின் சிக்கல்களை மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்டு உட்படுத்தப்பட்ட ஒரு மொழியின் அகராதியிலும் இலக்கணத்திலும் சேகரித்து வைக்கப்பட்டுள்ள ஏராளமான அறிவின் உதவியால் தீர்க்க இயலும். இது மொழிபெயர்ப்பு ஆய்வின் கடுமையான செயற்பாங்கின் பயன்படுத்தலை வேண்டும். சுருக்கமாக இயந்திர மொழிபெயர்ப்பில் மொழி ஆய்வு நுட்பத்திற்குப் பல பயன்பாடுகள் இருக்கின்றன. அவை பின்வருமாறு அமையும்:

- மொழிப்பெயர்ப்பு பகுப்பாய்வு முன்னரே இருக்கக் கூடிய மொழி பெயர்ப்புகளை அமைப்புகளை அமைப்பாக்கம் செய்ய உதவும். இதன் மூலம் அவற்றை புதிய மொழி பெயர்ப்புகளின் உருவாக்கத்தில் மறு உபயோகம் செய்ய இயலும்.
- மொழிப்பெயர்ப்பு பகுப்பாய்வு தொழில் நுட்பம் மொழிப்பெயர்ப்பு தவறுகளின் சில வகைகளைக் கண்டுபிடிக்க வேண்டி தொடக்கநிலை மொழிபெயர்ப்புகளில் பயன்படுத்தப்படுகிறது.
- மொழிப்பெயர்ப்பு பகுப்பாய்வு நுட்பம் இணை மொழிச் சொற்களால் ஏற்படும் குறுக்கீட்டுத் தவறுகளிலிருந்து சுதந்திரமாக மொழிபெயர்ப்பு இருந்தால் அதைப் பரிசோதிக்கப் பயன்படுத்தப்படும்.

10.11 இருமொழிய அகராதியின் உருவாக்கம்

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பின் மற்றொரு பயனுள்ள மற்றும் முக்கியப் பகுதி இருமொழி அகராதிகளின் உருவாக்கமாகும். இதன் குறைபாடு இந்திய மொழிகளின் இயந்திர மொழிபெயர்ப்பின் தடைகளில் ஒன்றாகும். மரபு அகராதியில் இக்குறைபாட்டை நிவர்த்தி செய்ய இயலாது. ஏனென்றால் அவை சொல் துணை வகைப்பாடு, சொல்தெரிவு, சொல்தேர்வுக் கட்டுபாட்டு மற்றும் சொல் அலகுகளின் பயன்பாட்டு களங்கள் இவற்றைப் பற்றிய போதுமான தகவல்களைக் கொண்டிருக்கவில்லை. இருமொழிய அகராதி உருவாக்கத்திற்கு நாம் பல வழிகளில் அடையாளப்படுத்தப்பட்ட மற்றும் பகுத்துக் குறிக்கப்பட்ட தரவுத்தொகுதிகளில் பயன்படுத்தப்பட்டுள்ள பல்வேறு புள்ளியியல் நெறிமுறைகளைப் பயன்படுத்த வேண்டும்.

10.12 மொழிபெயர்ப்பு நிகரன்களின் பிரித்தெடுப்பு

மொழிப்பெயர்ப்பு தரவுத்தொகுதிகளில் மொழிப்பெயர்ப்பு நிகரன்களைத் தேடல் மூல மொழியில் ஒரு குறிப்பிட்ட பொருண்மையை அல்லது கருத்துருவை வெளிப்படுத்தும் ஒரு குறிப்பிட்ட வடிவிலிருந்து தொடங்கப்பட வேண்டும். மூல

மொழியில் தேடல் செயற்பாங்கின் 2-வது பகுதி தொடங்கும். இது இலக்குமொழியில் ஒரே அர்த்தத்தைக் கொண்டிருக்கிற சொல்லின் அடையாளம் காணலை உட்படுத்தும். இயல்பாகப் பெரும்பாலான மொழிபெயர்ப்பு தரவுத்தொகுதிகள் மொழிபெயர்ப்பு நிகரன்களின் மிகுந்த பரப்பெல்லையை வெளிப்படுத்த இயலும். இவை மாற்று வடிவின் தேர்வை உறுதி செய்யும் காரணிகள், சொற்கள் பயன்பாட்டின் மறு நிகழ்வு அமைப்பொழுங்கின் அடிப்படையில் உறுதி செய்யப்படும்.

10.13 கலைச்சொல் தகவல் வங்கியின் உருவாக்கம்

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிப்பெயர்ப்பு தொழில் நுட்ப மற்றும் அறிவியல் கலைச்சொற்களில் பொருத்தமான ஆக்கத்திற்குக் கலைச்சொல் ஆக்கவியலாரின் எண்ணத்தின் அடிப்படையிலான ஆய்வு தேவை.கலைச்சொல் ஆக்கவியலாரின் அடிப்பசைட் செயல்பாடு இலக்கு மொழிக்கு அயல் கருத்துக்களையும் கருத்துருக்களையும் உருப்படுத்தம் செய்ய மிகப் பொருத்தமான அல்லது ஒரளவுக்குப் பொருத்தமான சொற்களைத் தேர்வு செய்வதாகும். இதைச் செய்வதற்குக் கலைச்சொல் ஆக்கவியலார் இலக்கு மொழி பயன்படுத்துபவர்களுக்கிடையில் கலைச்சொற்களின் பொருத்தமான தன்மை, இலக்கணத் தன்மைம, ஏற்றுக்கொள்கை மற்றும் பயன்படு பல்வேறு காரணிகளை மனதில் கொள்ளவேண்டும். முக்கியமான காரணி கலைச்சொற்களின் தரம் ஆகும். இதன்படி மொழியியல் கொள்கைகளின் அடிப்படையில் பொருத்தமான வழிகளைப் பயன்படுத்தி புதிய சொற்கள் உருவாக்கப்டும். மொழிபெயர்ப்பு தரவுத்தொகுதி ஒரு கருத்து, நிகழ்ச்சி, விக்ஷயம் அல்லது கருத்துரு என்பதை உருப்படுத்தம் செய்யப் பலரால் உருவாக்கப்பட்ட பல கலைச்சொற்களின் பெரிய பட்டிலிலிருந்து பொருத்மான ஒன்றைத் தேர்வு செய்யும் முக்கிய பங்களிப்பைச் செய்கின்றது. புதிய சொற்கள் உருவாக்கத்தின் தொடர்ச்சியான முயற்சி தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பு அமைப்பைத் திட்டமிடுவர்களுக்கு ஒன்றுக்குப் பதிலாக ஒன்றைத் தேர்ந்தெடுப்பதில் சிக்கலைத் தரும். புதியச் சொல்லை உருவாக்குவதா அல்லது இலக்கு மொழியில் பயன்பாட்டால் இயல்பாக்கம் செய்யப்பட்ட

மூலமொழிச் சொல்லைப் பயன்படுத்துவதா என்ற விவாதம் எழும். சுப்ல தொழில் நுட்பச் சொற்கள் அவற்றின் உண்மையான மூலத்தைக் கண்டுபிடிக்க இயலாதவாறு இயல்பாக்கம் செய்யப்பட்டிருக்கும். அன்றாடம் வாழ்க்கையில் அவற்றைப் பயன்படுத்தும் உயர்ந்த நிகழ்வெண் அவற்றைச் சொற்றொகையின் பகுதியாக மாற்றும். எனவே மொழிபெயர் பின் போது அவற்றை இடம் பெயர்க்கத் தேவையில்லை. பொதுவாக இலக்கு மொழியில் பெயர்க்கப்பட்ட தரவுத்தொகுதி மூல மொழியிலிருந்து கடன் வாங்கப்பட்டு புதிய கருத்துக்களையும் கருத்துருக்களையும் வெளிப்ப்படுத்தப் பொருத்தமான சொற்களைத் தேர்ந்தெடுக்க நல்ல மூலவளமாகும்.

10.14 சொல் தேர்வு கட்டுப்பாடு

மூல மொழியிலிருந்து தெரிந்தெடுக்கப்பட்ட மிகப் பொருத்தமான நிகரன்காளகக் கருதப்பட வேண்டிய இலக்கு மொழிக்கு சொற்களின் தெரிவு மூலமொழியிலும் பட்டறிவுள்ள மொழியியலாரின் திறமையான மற்றும் கவனமான பொரள்கோளை வேண்டும் மற்றொரு கலவைத்தன்மையான செயல்பாடாகும். சூதன் பொருள் என்னவென்றால் மொழியியலார் இலக்கு மொழியில் இருக்கின்ற ஒரே பொருண்மையத் தருகின்ற வடிவங்களின் பெரிய தொகுப்பிலிருந்து கவனத்தில் கொள்ளவும்.

ஆங்கிலம் : that saint ate food
தமிழ் : அந்த முனிவர் போஜனம் செய்தார்
ஆங்கிலம் : he ate food
தமிழ் : அவன் உணவு சாப்பிட்டான்

மேற்சொன்ன எடுத்துக்காட்டுகளில் நாம் தமிழில் எழுவரின் பண்பின் அடிப்படையில் சரியான இணைச் சொல்லைக் தெரிந்தெடுக்க வேண்டியுள்ளது. மொழியியலாரின் முக்கியமான செயல்பாடு இரு மொழிகளுக்கு இடையிலான மொழிபெயர்ப்பில் கருத்தப்படும் வேறுபட்ட சமூக மொழியியல் காரணிகளுக்குப் பொருத்தமான சொல்

அலகுகளைக் காண்பதாகும். இம்மாதிரியான எடுத்துக்காட்டுகள் வெற்றிகரமான மற்றும் அறிவு பூர்வமான மொழிபெயர்ப்பில் சொல் தேர்வு கவனமாகக் கையாளப்பட வேண்டும் என்பதைக் காட்டப்படுகினறது. இச்சிக்கல் மனித மொழிபெயர்ப்பில் கவனமாகக் கையாளப்பட்டாலும் பெரும்பாலான இயந்திர மொழிபெயர்ப்பில் புறக்கணிக்கப்படுகினறன. இச்சிக்கலைத் தீர்ப்பதற்கான நல்ல வழி இயந்திரத்தால் படிக்க இயலும் அகராதியின் தனியான இடத்தில் பொருண்மையில் ஒற்றுமையைக் காட்டும் எல்லா வடிவங்களையும் பட்டியிலிடுவதாகும். இம்மாதிரியான தரவுமையங்களை மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பிரித்தெடுக்க இயலும், பொதுவாக இயந்திரம் படிக்க இயலும் அகராதிகள் பல பாடப்பொருள் களங்களாகப் பிரிக்கப்பட்டிருக்கும். இதனால் ஒரு குறிப்பிட்ட பாடப்பொருளைச் சார்ந்த சொற்களைப் பற்றிய பொருத்தமான தகவலைப் பெற இயலும்.

10.15 சொல் மயக்கத்தை நீக்குதல்

இயல்பான சூழலில் மனித கருத்துப் பரிமாற்றம் ஒழுங்கமைப்பு மொழியைக் கருவியாக கொண்டு பேசுபவரிடமிருந்து கேட்பவருக்குச் செய்தியைப் பரிமாற்றம் செய்வதை உள்ளடக்கும். சில நேரங்களில் செய்தியின் பரிமாற்றம் பொருண்மை மயக்கத்திலிருந்து விடுபடாமல் இருக்கும். இப்பொருண்மை மயக்கம் மொழியில் மிகப் பரவலான நடப்பாகும். மயக்கமான சொல்லுடன் தொடர்புடைய அகப்பொருண்மையின் போதாமையால் அல்லது கருத்துப் பரிமாற்றத்தின் ஒரு குறிப்பிட்ட நிகழ்வில் கூற்றின் அமைப்பால் ஏற்படும். இவ்வாறு மயக்கங்கள் சொல் மயக்கம் (lexical ambiguity) என்றும் குறிப்புமரை மயக்கம் (Referential ambiguity) என்றும் தொடரியல் மயக்கம் (sytacitic mbiguity) என்றும் மூன்றாகப் பகுக்கப்படும்.

they went to bank	-	சொல் மயக்கம்
he love his wife	-	குறிப்புரை மயக்கம்
he waw a girl in part with a binocular	-	தொடரியல் மயக்கம்

இயந்திர மொழிபெயர்ப்பு ஒங்கமைப்பு பேசுபவரின் மன உருப்படுத்தத்தின் புலனுணர்வு அடிப்படையில் உருவாக்கப்படுவதால் இது பேசுபவரால் பயன்படுத்தப்படும் சொற்களுக்கும் வாக்கியங்களுக்கும் எல்லைப்படுத்தப் பட்டிருக்கும். இதை நேர் செய்ய மொழிபெயர்ப்பாளர்கள் மூல அகராதியை ஒரு குறிப்பிட்ட சூழலில் பொருத்தமாக் தோன்றுகின்ற நிகரனை இலக்கு அகராதியால் பொருத்துகிறார்கள். சில நேர்வுகளில் இலக்கு மொழியானது மூலமொழியின் சொல்லை உருப்படுத்தம் செய்ய இயலும் நிகரான சொல் அலகைக் கொண்டிருக்காது. இந்நேர்வுகளில் மொழிபெயர்ப்பாளர்கள்ன சொற்றொடரைப் பயன்படுத்தப் முயல்வர் அல்லது நிகரான விளக்கத்தைத் தர முயல்வர்.

தரவுத்தொகுதி அடிப்படையிலான இயந்திர மொழிபெயர்ப்பை ஆதரிப்பவர்கள் சொல் மயக்கச் சிக்கலை நீக்க இயற்க்கைமொழி உரைகளில் கூடுதலாக நிகழும் மயக்க வடிவுகளில் பெரும் எண்ணிக்கையை ஆய்ந்தும் வெளிப்ப்படையாக உருப்படுத்தம் செய்தும் உயர்ந்த தரமான மொழிபெயர்ப்பை தரலாம் என்று வாதியிடுகின்றனர். சத்தியம் என்றால் மயக்கமான வடிவுகளும் மொழிபெயர்ப்பு தரவுத்தொகுதிகளிலிருந்து பெறப்பட்ட விரிந்த அறிவின் அடிப்படையில் ஆயப்பட்டு இயந்திரத்தால் புரிந்து கொள்ளப்படும் வடிவுகளும் அகராதியில் சேமித்து வைக்கப்பட்ட வேண்டும். ஆனால் பகுத்தறிவாதிகள் இவ்விதம் மொழியியல் தகவல்களைப் பெறும் செயல்பாடு சரியானதும் அல்ல சாத்தியமானதும் அல்ல என்று வாதியிடுகின்றனர். களக்கட்டுப்பாடு மொழிபெயர்ப்பு (domin – specific translation) மொழிபெயர்ப்புக்கு எப்போதும் ஆழமான பொருண்மையியல் ஆய்வு தேவையில்லை என்று வெளிப்ப்படுத்துகின்றன. குறிப்பிட்ட மயக்கம் உள்ள சொல் அலகுகளின் ஆழமான பொருண்மையியல் ஆய்வுக்குப் பதிலாக நேரடியாக மூல மொழியிலிருந்து இலக்கு மொழிக்குப் பதிலீடு செய்யும் அணுகுமுறையான எளிய ஆய்வைத் தெரிந்தெடுப்பது நல்லது. சில சூழல்களில் மயக்கங்கள் மொழி பெயர்ப்பின் போது அப்படியே கொண்டு செல்லப்படும் என்ற எதிர்ப்பு அடிப்படையில் சில சொல் மயக்கங்களைப் புறக்கணித்தல் சாத்தியமானதும் தேவையானதுமாகும். இருப்பினும் சொல் மயக்கங்களின் ஆய்வு, இலக்கு மொழியில்

மயக்கமற்ற உருப்படுத்தங்களை உருவாக்கும் என்ற நிலையில் பொதுவான மொழிபெயர்ப்பு நேர்வுகளில் மயக்கத்தைப் புறக்கணிக்கக் கூடாது.

10.16 இலக்கணப் பொருத்தம்

மொழியில் மொழிபெயர்ப்பில் இலக்கணப் பொருத்தம் என்பது மூல மொழியில் ஒரு சொல் இலக்கு மொழியில் உள்ள சொல்லுடன் அர்தம் உள் மொழிபெயர்ப்பைப் பெற வேண்டி பொருத்தப்படும் மொழிபெயர்ப்பில் பொருத்தத்திற்கு வேறுபட்ட திட்டங்கள் இருக்கின்றன. அவை சொல் நிலைப் பொருத்தம், உருபன் நிலைப் பொருத்தம், இலக்கணப் பொருத்தம், தொடர்நிலைப் பொருத்தம், வாக்கியநிலைப் பொருத்தம் என்பனவாகும். மிகப் பொதுவானது மொழிபெயர்ப்புக்கு எடுத்துக்கொள்ளப்ட்ட இரு மொழிகளுக்கிடையில் வினை பங்களிப்பைப் பொருத்துவதாகும. இலக்கண பொருத்தம் வகைப்பாட்டியியல் அடிப்படையில் வேறுபட்ட இரண்டு மொழிகளுக்கிடையில் செய்யப்படும் இயந்திர மொழி பெயர்ப்பில் முக்கியத்துவம் உள்ளதாகும். தற்போதைய சூழலில் நாம் ஆங்கிலத்திலிருந்து தமிழுக்கு மொழிபெயர்ப்பைப் பற்றி பேசும் பொது இது முக்கியத்துவம் வாய்ந்தது: என்னவென்றால் ஆங்கிலத்திற்கு SVO அமைப்பு உண்டு He is food தமிழுக்கு SOV அமைப்பு இருக்கிறது (அவன் சோறு சாப்பிட்டான்). எனவே இலக்கணப் பொருத்தமும் சொல் வரிசைமுறை மாற்றமும் இலக்கு மொழியில் வெளிப்யீடுகளை வெளிப்டுவதற்கு முக்கியத்துவம் வாய்ந்ததாகும்.

11. முடிவுரை

முந்தைய பகுதிகளில் தரவுத்தொகுதியின் விளக்கம், தரவுத்தொகுதியின் தொடக்க நிலை, தரவுத்தொகுதியின் இன்றைய நிலை, இதுவரை உருவாக்கப்பட்ட தரவுத்தொகுதிகள், தரவுத்தொகுதியின் வகைப்பாடு, எழுத்துவடிவத் தரவுத்தொகுதியின் உருவாக்கம், உரை தரவுத்தொகுதி ஆய்வு, மொழித் தொழில் நுட்பத்தில் தரவுத்தொகுதி, முக்கிய மொழியியலில் தரவுத்தொகுதி, இயந்திர மொழிபெயர்ப்பில் தரவுத்தொகுதி, தரவுத்தொகுதியின் பயன்பாடுகள் என்ற தலைப்புகளில் தரவுத்தொகுதி தொடர்பான செய்திகள் எல்லாம் விளக்கமாக எடுத்துரைக்கப்பட்டுள்ளன. களமொழியலிருந்து

தரவுத்தொகுதி மொழியலுக்கான பயணம் பல மயில்கற்களைத் தாண்டி வந்துள்ளது. அன்றைய கால கட்டத்தில் மொழியியலார் கள ஆய்வை மேற்கொண்டு மொழித் தரவுகளைத் திரட்டி ஒரு மொழியின் அமைப்பை அறிய ஒலியியல், ஒலியனியல், உருபனியல், தொடரியல் மற்றும் பொருண்மையியல் மட்டங்களில் ஆய்வு செய்வார்கள். இத்தகைய ஆய்வு தொடக்க காலத்தில் அமைப்புமுறை மொழியியல் ஆய்வாக அமைந்தது. இது நடத்தையிலாரால் ஊக்கப்படுத்தப்பட்ட ஊன்றிக்கவனித்தல் அடிப்படையிலான அனுபவவாத அடிப்படையில் நிகழ்த்தப்பட்டது. சாம்ஸ்கி இத்தகைய ஆய்வுகளைக் செயல்திறன் சார்ந்தது என்றும் குறையுள்ளது என்றும் கூறிப் பகுத்தறிவு வாதத்திற்கு வித்திட்டார். இதன்பயனாகக் கள ஆய்வுகள் ஊக்கமிழந்து உள்ளுணர்வு ஆய்வுகள் ஊக்கமடைந்தன. சாம்ஸ்கி கள ஆய்வாளர்களால் திரட்டப்பட்ட தரவு குறையுடையது என்று வாதிட்டார். ஆனால் தரவுத்தொகுதியின் வரவால் சாம்ஸ்கி கூறிய குறைபாடுகள் நிவர்த்தி செய்யப்பட்டு மொழியில் ஆய்வுகள் கணிமொழியியல் ஆய்வாக மாறியது. உள்ளுணர்வு அடிப்படையிலான பல கருதுகோள்கள் தவறானவை என்பதை தரவுத்தொகுதி ஆய்வு நமக்கு உணர்த்தியது. தரவுத்தொகுதிகள் மொழி வல்லுனர்களால் திறமையாகப் பயன்படுத்தப்படுகின்றது. அகராதிவியலார்கள், கலைச்சொல்லாக்க வல்லுனர்கள், கோட்பாட்டியலார்கள், பயன்பாட்டு மொழியலார்கள் தத்தம் செயல்பாடுகளுக்கு இன்றைய காலகட்டத்தில் தரவுத்தொகுதி பெரிதும் பயன்படுத்துகின்றர். தரவுத்தொகுதியைக் கொண்டு அகராதிகள் உருவாக்கப்படுகின்றன. கலைச்சொல்லாக்கத்திற்காகத் தரவுத்தொகுதி பெரிதும் உதவுகின்றது. கோட்பாட்டாளர்கள் தங்கள் கோட்பாடுகளை நிரூபிப்பதற்கும் தங்கள் கருதுகோள்களின் உண்மை நிலையை அறியவும் தரவுதொகுதியைப் பயன்படுத்துகின்றனர். பயன்பாட்டு மொழியிலார் பல மொழியில் பயன்பாடுகளுக்குத் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். கணினி மொழியிலார்கள், கற்றல் கற்பித்தல் வல்லுனர்கள், சமூக மொழியலார்கள், உளமொழியியலார்கள், மானிட மொழியலார்கள், பண்பாட்டு மொழியியலார்கள் போன்றோர் தரவுத்தொகுதியைத் தங்கள் ஆய்விற்கும் பயன்பாட்டுக் கருவிகள் தயாரிக்கவும் பயன்படுத்துகின்றனர். பொருளடக்க வல்லுனர்களாக வரலாற்றியலார்கள், இலக்கியத் திறனாய்வாளர்கள், சமூகவியலார்கள் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். வரலாற்றியலார்களுக்குத் தரவுத்தொகுதி பல வழிகளில் கைகொடுக்கின்றது. தரவுத்தொகுதியைப் பயன்படுத்தி ஒரு நிகழ்வின் வரலாற்று காலம், இடம் போன்ற கூறுகளைச் சரியாகக் கணிக்கின்றனர். சமூகவியலார்

வேறுபட்ட வகுப்பு, இனம், குடி சார்ந்த உண்மைகளையும் தகவல்களையும் தரவுத்தொகுதியின் உதவியால் சரியாகக் கணிக்கின்றனர். தகவல் வல்லுனர்களாகத் தகவல் மீட்பாளர்கள், கணினி மொழியியலார்கள், இயந்திர மொழிபெயர்ப்பாளர்கள், மொழியாய்வு ஆய்வாளர்கள் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். தகவல் மீட்பாளர்கள் தரவுத்தொகுதியிலிருந்து தேவையான தகவல்களைப் பெற வேண்டி பல நுட்பங்களை உருவாக்கிச் செயல்படுகின்றனர். அவர்கள் அறிவு அடிப்படை உருவாக்கத்திற்கும் செயற்கை அறிவுக்கும் தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். கணினி மொழியியலார்கள் மொழியை ஆயவும் பகுப்பாயவும் முக்கியமான உபாயமாகச் செயல்படும் தரவுத்தொகுதியில் காணப்படும் புள்ளியியல் முறைமையுடன் தங்கள் செயல்பாட்டை ஒருங்கிணைக்க தரவுத்தொகுதியைப் பயன்படுத்துகின்றனர். மேலும் தரவுத்தொகுதி தரவு மற்றும் அறிவடிப்படையின் மூலவளமாக, புள்ளியியல் உத்திகள் இலக்கண அடிப்படையில் ஆயப்பட்ட தரவுத்தொகுதிகளின் விடுவரல்களின் மீது செயல்படுகையில் கூடுதல் திறனுள்ளதாய் மாறுவதன் காரணமாக மொழியில் உள்ள முறைமைகளின் இருப்பை அல்லது இல்லாமையைப் பரிசோதிக்கப் பயன்படுத்தப்படுகின்றது. இயந்திர மொழிபெயர்ப்பாளர்கள் தரவுத்தொகுதி ஒழுங்குமுறைகளின் உண்மையான திறனை அதிகரிக்க வேண்டி குறிப்பிடத்தக்கப் பங்களிப்பு செய்வதன் காரணமாகத் தேவையான மற்றும் பொருத்தமான மொழியியல் தகவல்களைப் பெறவும் அவர்களின் ஒழுங்குமுறைகளின் திறமையைப் பரிசோதிக்கவும் தரவுத்தொகுதியை நாடுகின்றனர். மொழி ஆய்விகளை உருவாக்கவேண்டி குறிப்புரை செய்யப்படாத மற்றும் குறிப்புரை செய்யப்பட்ட தரவுத்தொகுதிகளைக் கூடுதலாகப் பயன்படுத்தப்படுவதால் மொழி ஆய்வாளர்கள் பல்வேறு வகைப்பட்ட தரவுத்தொகுதிகளின் உருவாக்கத்தால் நன்மை அடைகின்றார்கள். தரவுத்தொகுதி ஆய்வாளர்கள், தொழிநுட்பவியலார்கள், எழுத்தாளர்கள், அகராதியலார்கள், கல்வியாளர்கள், ஆசிரியர்கள், மாணவர்கள், மொழிகற்பவர்கள், அறிவாளிகள், வெளியீட்டாளர்கள் மற்றும் பிறர் உள்ளடங்கிய எல்லோருக்கும் பயனுள்ளதாக இருக்கின்றது. இவ்வாறு தரவுத்தொகுதி இயந்திர மொழிபெயர்ப்புக்குப் பெரிதும் பயனுள்ளதாக அமைகிறது. இணையான மொழிபெயர்க்கப்பட்ட தரவுத்தொகுதி சொல் பொருத்தம், உருபன் பொருத்தம், இலக்கணப் பொருத்தம், தொடர் பொருத்தம், வாக்கியப் பொருத்தம் என்ற செயல்பாடுகளிலிருந்து மொழியின் சொல் மட்டம், உருபன் மட்டம்,

தொடரியல் மட்டம் என்ற நிலைகளில் நிகரன்களைக் கண்டறிந்து கூடுதலாக தரவுத்தொகுதிப் பயிற்சி தந்து இயந்திர மொழிபெயர்ப்பை மேம்படுத்தலாம்.

Aarts, J. and Meijs, W. (Eds.) 1984. *Corpus Linguistics: Recent Development in the Use of Computer Corpora in English Language Research*. Amsterdam-Atlanta, GA.: Rodopi.

Aarts, J. and Meijs, W. (Eds.) 1986. *Corpus Linguistics II: New Studies in the Analysis and Explanation of Computer Corpora.* Amsterdam-Atlanta, GA.: Rodopi.

Aijmer, K. and Allwood, J. (Eds.) 2004. *Dialogic Analysis: New Trends in Dialogue A*nalysis. Tubingen: Niemeyer.

Aijmer, K. and Altenberg, B. (Eds.) 1991. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman.

Altenberg B. and Granger, S. (Eds.) 2002. *Lexis in Contrast: Corpus-based Approaches*. Amsterdam/ Philadelphia: John Benjamins.

Aston, G. (Ed.) 2004. *Learning with Corpora*. Cambridge: Cambridge University Press.

Atwell, E. (Eds.) 1993. *Corpus Based Computational Linguistics*. Amsterdam: Rodopi.

Baker, M., Gill, F. and Tognini-Bonelli, E. (Eds.) 1993. *Text and Technology: In honour of John Sinclair.* Philadelphia: John Benjamins.

Barnbrook, G. 1998. *Language and Computers*. Edinburgh: Edinburgh University Press.

Béjoint, H. 2000. *Modern Lexicography: An Introduction*. Oxford: Oxford University Press.

Bernadini, S. 2000. *Competence, Capacity, Corpora*. Bologna: CLUEB.

Bernini G. (Ed.) 2001. *Pragmatic Organisation of Discourse in the languages of Europe*. Vol. 1. Berlin: Mouton de Gruyter.

Bharati, A., Chaitanya, V. and Sangal, R.1995. *Natural Language Processing: A Paninian Perspective*. New Delhi: Prentice-Hall.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, D., Conrad, S. and Reppen, R. 1998. *Corpus linguistics - Investigating Language Structure and Use*. Cambridge: Cambridge University Press.

Black, A.H. and Black, C.A. 1997. *A Conceptual Introduction to Morphological Parsing Using AMPLE System*. Dallas, Texas: Summer Institute of Linguistics.

Black, E., Garside, R. and Leech, G. (Eds.) 1993. *Statistically-driven Computer Grammars of English: the IBM/Lancaster Approach*. Amsterdam: Rodopi.

Bloom, L. 1970. *Language Development: Form and Function in Emerging Grammars*. Cambridge, Mass.: MIT Press.

Boguraev, B. and Pustejvsky, J. (Eds.) 1996. *Corpus Processing for Lexical Acquisition*. Cambridge, Mass.: MIT Press.

Botley, S.P., A.M. Mcenery, and A. Wilson (Eds.) 2000. *Multilingual Corpora in Teaching and Research*. Amsterdam-Atlanta, GA: Rodopi.

Bouillon, P. and Busa, F. (Eds.) 2001. *The Language of Word Meaning*. Cambridge: Cambridge University Press.

Brown, R. 1973. *A First Language: The Early Stages*. Cambridge, Mass.: Harvard University Press.

Burton, D.M. 1973. *Shakespeare's grammatical style: a computer-assisted analysis of Richard II and Anthony and Cleopatra*. Austin: University of Texas Press.

Butler, C.S. (Ed.) 1992. *Computers and Written Texts*. Oxford: Blackwell Publishers.

Carter, R. and Mccarthy, M. (Eds.) 1988. *Vocabulary and Language Teaching*. London: Longman.

Cassidy, F.G. Ed. 1985. *Dictionary of American Regional English*. Vol.1: Introduction and A-C. Cambridge, MA.: Harvard University Press.

Chomsky, A.N. 1968. *Language and Mind*. New York: Harcourt Brace.

Chomsky, A.N. 1972. *Studies in Generative Grammar*. The Hague: Mouton.

Coleman, J. and Kay, C.J. (Eds.) 2000. *Lexicology, Semantics and Lexicography: Selected Papers from the 4th G.L. Brook Symposium.* Amsterdam/ Philadelphia: John Benjamins.

Collins, P. and Blair, D. (Eds.) 1989. *Australian English*. St. Lucia: University of Queensland Press.

Crystal, D. 1995. *The Cambridge Encyclopaedia of the English Language*. Cambridge: Cambridge University Press.

Cuyckens, H. and ZAWADA, B. (Eds.) 2001. *Polysemy in Cognitive Linguistics.* Amsterdam/Philadelphia: John Benjamins.

Dash, N.S. 2005. *Corpus Linguistics and Language Technology (With Reference to Indian languages).* New Delhi: Mittal Publications.

Desai, S.K. 1974. *Experimentation with language in Indian Writing in English (Fiction).* Monograph of the Dept. of English, Shivaji University, Kohlapur, India.

Edwards, J.A. and Lampert, M.D. (Eds.) 1993. *Talking Data: Transcription and Coding in Discourse Research.* Hillsdale, N.J.: Lawrence Erlbaum Associates.

Eggins, S. 1994. *An Introduction to Systemic Functional Linguistics*. London: Pinter Publishers.

Flesch, R. 1946. *The Art of Plain Talk.* New York: Harper & Brothers.

Flowerdew, J. and Peacock, M. (Eds.) 2001. *Research Perspectives on English for Academic Purposes*. Cambridge: Cambridge University Press.

Francis, W.N. and Kucera, H. 1964. *Manual of information to accompany A standard Corpus of present-day edited American English.* Dept. of Linguistics, Brown University, USA.

Francis, W.N. and Kucera, H. 1982. *Frequency Analysis of English Usage: Lexicon and Grammar.* Boston: Houghton Mifflin.

Fries, U., Mü, V. and Schneider, P. (Eds.) 1997. *From Aelfric to the New York Times*. Amsterdam: Rodopi.

Fuchs, C. and Vitorri, B. (Eds.) 1994. *Continuity in Linguistic Semantics*. Amsterdam and Philadelphia: John Benjamins.

Garside, R., Leech, G. and Mcenery, A. (Eds.) 1997. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Garside, R., Leech, G. and Sampson, G. (Eds.) 1987. *The Computational Analysis of English: A Corpus Based Approach.* London: Longman.

Gerbig, A. 1997. *Lexical and Grammatical Variation in a Corpus: A Computer-Assisted Study of Discourse on the Environment.* London: Peter Lang Publishing.

Ghadessy, M., Henry, A. and Roseberry, R.L. Eds. 2001. *Small Corpus Studies and ELT: Theory and Practice.* Amsterdam/ Philadelphia: John Benjamins.

Gibson, H.N. 1962. *The Shakespeare Claimants: A Critical Survey of the Four Principle Theories Concerning the Authorship of the Shakespearean Play*. London: Methuen and Co.

Gove, P.B. (Ed.) 1961. *Webster's Third New International Dictionary of the English Language Sentences*. New York: Hartcourt, Brace and World.

Graf, D. 1996. *Relative Clauses in Their Discourse Context: A Corpus-Based Study*. Unpublished M.A. Thesis: Freiburg.

Granger, S. and Tyson, S.P. (Eds.) 2003. *Extending the Scope of Corpus-Based Research: New Applications, New Challenges*. Amsterdam: Rodopi.

Granger, S., Hung, J. and Tyson, S.P. Eds. 2002. *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching*. Amsterdam: John Benjamins.

Greenbaum, S. (Ed.) 1996. *Comparing English Worldwide: The International Corpus of English.* Oxford: Clarendon.

Greene, B. and Rubin, G. 1971. *Automatic Grammatical Tagging of English.* Technical Report. Department of Linguistics. Brown University, RI, USA.

Halliday, M.A.K. 1987. *Spoken and Written Modes of Meaning, Comprehending Oral and Written Language.* San Diego, CA: Academic Press.

Halliday, M.A.K. 1989. *Spoken and Written Language*. Oxford: Oxford University Press.

Halliday, M.A.K. and Hasan, R. 1976. *Cohesion in English*. London: Longman.

Halteren, H.V. (Ed.) 1999. *Syntactic Word Class Tagging*. Dordrecht: Kluwer Academic Press.

Hausmann, F.J., Reichmann, O., Wiegand, H.E. and Zgusta, L. (Eds.) 1990. *International Encyclopedia of Lexicography.* Vol.2. Berlin: Walter de Gruyter.

Herden, G. 1962. *Calculus of Linguistic Observation*. Hague: Mouton & Co.

Hickey, R. and Stanislaw, P. (Eds.) 1996. *Language History and Linguistic Modelling: A Festschrift for Jacek Fisiak.* Vol. 2. Berlin: Mouton de Gruyter.

Hoffman, C. 1955. *The Man Who was Shakespeare*. New York: Julias Messner Inc.

Hofland, K. and JOHANSSON, S. 1982. *Word Frequencies in British and American* English. Bergen: Norway Computing Centre for the Humanities.

Hunston, S. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Hutchins, W. J. 1986. *Machine Translation: Past, Present, and Future.* Chichester: Ellis Harwood.

Ide, Nancy and Jean Véronis (Eds.) 1998. *Introduction to the special issue on word sense disambiguation: the state of the art*. Computational Linguistics. 24(1): 1-40.

Ilson, R.F. (Ed.) 1986. *Lexicography: An Emerging International Profession.* Manchester: Manchester University Press.

Ingram, D. 1989. *First Language Acquisition*. Cambridge: Cambridge University Press.

Jensen, J.T. 1990. *Morphology: Word Structure in Generative Grammar*. Amsterdam: John Benjamins.

Jespersen, O. 1909-1949. *Modern English Grammar on Historical Principles*. 7 Vols. London: Allen and Unwin.

Johansson, S. and Hofland, K. (Eds.) 1982. *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Johansson, S. and Stenström, A-B. (Eds.) 1991. *English Computer Corpora: Selected Papers and Research Guide*. Berlin: Mouton de Gruyter.

Katamba, F. 1993. *Morphology*. London: Macmillan Press.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics.* New York: Addison-Wesley Longman Inc.

Kenny, A.J.P. 1982. *The Computation of Style*. Oxford: Pergamon Press.

Kettemann, C.B. and Marko, G. (Eds.) 2002. *Teaching and Learning by Doing Corpus Analysis. Language and Computers:* Studies in Practical Linguistics 42. Amsterdam-Atlanta, GA.: Rodopi.

Kilgariff, Adam and J. Palmer (Eds.) 2000. *Computer and the Humanities: Special Issue on Word Sense Disambiguation.* Vol. 34. No.1. 2000.

Kirk, J.M. (Ed.) 2000. *Corpora Galore: Analyses and Techniques in Describing English.* Amsterdam; Atlanta, GA: Rodopi.

Kruisinga, E. 1931-32. *A Handbook of Present-Day English*. Goringen: Noordhoff.

Kucera, H. and Francis, W.N. 1967. *Computational Analysis of Present Day American English*. Providence, RI: Brown University Press.

Kurath, H. Ed. 1954. *Handbook of the linguistic geography of New England*. Washington DC: American Council of Learned Societies.

Kytö, M., Ihalainen, O. and Rissanen, M. (Eds.) 1988. *Corpus Linguistics, hard and soft: Proceedings of the 8th International Conference on English Language Research on Computerised Corpora.* Amsterdam: Rodopi.

Lancashire, I., E. Carol, and C.F. Meyer (Eds.) 1997. *Synchronic Corpus Linguistics*. Bergen, Norway: ICAME.

Landau, S.I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Cambridge: Cambridge University Press.

Leech, G., G. Myers, and J. Thomas (Eds.) 1995. *Spoken English on Computer: Transcription, Markup and Applications*. Harlow: Longman.

Levy, M. 1997. *Computer Assisted Language Learning.* Oxford: Oxford University Press.

Ljung, M. (Ed.) 1997. Corpus-Based Studies in English. Papers from the 17th International Conference on English-Language Research Based on Computerized Corpora. Amsterdam: Rodopi.

Lock, A. (Ed.) 1978. *Action, Gesture and Symbol: The Emergence of Language*. London: Academic Press.

Lorge, I. 1949. *Semantic Content of the 570 Commonest English Words*. New York: Columbia University Press.

Macwhinney, B. 1991. *The CHILDES Project: Tools for Analyzing Tal*k. Hillsdale, N.J.: Lawrence Erlbaum.

Mair, C. and Hundt, M. (Eds.) 2000. *Corpus Linguistics and Linguistics Theory*. Amsterdam-Atlanta, GA: Rodopi.

Mcarthur, T. 1981. *Longman Lexicon of Contemporary English*. London: Longman.

Mccarthy, J. 1982. *Formal Problems in Semitic Phonology and Morphology*. New York: Garland.

Mccarthy, M. 1998. *Spoken Language and Applied Linguistics*. Cambridge: cambridge University press.

Mcenery, T. and Wilson, A. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Mcenery, T., Rayson, P. and Wilson, A. (Eds.) 2002. *A Rainbow of Corpora: Corpus Linguistics and the Langauges of the World.* München: Lincom Europa.

Meyer, C.F. 2002. *English Corpus Linguistics*. Cambridge: Cambridge University Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K. 1990. *Five Papers on WordNet. CSL Report 43,* Cognitive Science Laboratory, Princeton University, Princeton.

Miller, G.A. 1951. *Language and Communication*. New York: McGraw-Hills.

Mindt, D. 1995. *An Empirical Grammar of the English Verb: Modal Verbs*. Berlin: Cornelsen Verlag.

Mitkov, R. (Ed.) 2003. The Oxford Handbook of Computational Linguistiucs. Oxford: Oxford University Press.

Nelson, G., WALLIS, S. and AARTS, B. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English.* Amsterdam/ Philadelphia: John Benjamins.

Nihalani, P., TONGUE, R.K. and HOSALI, P. 1979. *Indian and British English: A handbook of Usage and pronunciation.* New Delhi: Oxford University Press.

Oakes, M.P. 1998. *Statistics for Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Ooi, V.B.Y. 1997. *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press

Oostdijk, N. and Dehann, P. (Eds.) 1994. *Corpus Based Research into Language*. Amsterdam-Atlanta, GA: Rodopi.

Orton, H. 1962. *Introduction to Survey of English Dialects*. Leads: E.J. Arnold and Son Ltd.

Palmer, H. 1933. *Second Interim Report on English Collocations.* Tokyo: Institute for Research in English Teaching.

Partington, A. 1998. *Patterns and Meanings - Using Corpora for English Language Research and Teaching.* Amsterdam/Philadelphia: John Benjamins.

Percy, C., Meyer, C.F. and Lancashire, I. (Eds.) 1996. *Synchronic Corpus Linguistics.* Amsterdam-Atlanta, GA: Rodopi.

Peters, B.P., Collins, P. and Smith, A. (Eds.) 2002. *New Frontiers of Corpus Research*. Language and Computers. Amsterdam-Atlanta, GA: Rodopi.

Pinker, S. 1995. *The Language Instinct: The New Science of Language and Mind*. Middlesex, England: Penguin Books Ltd.

Poutsma, H. 1926-29. *A Grammar of Late Modern English*. Groningen: Noordhoff.

PREYER, W. 1989. *The Mind of a Child*. New York: Appleton.

Pustejovsky, J. 1995. *The Generative Lexicon*. Cambridge, MA: MIT Press.

Quirk, R., Gimson, A.C. and Warburg, J. 1968. *The Use of English*. 2nd Edition. London: Longman Group Ltd.

Quirk, R., Greenbaum, S., Leech, G. and Svartvik, J. 1985. *A Comprehensive Grammar of the English Language.* London: Longman.

Ravin, Y. and Leacock, C. (Eds.) 2000. *Ploysemy: Theoretical and Computational Approaches*. New York: Oxford University Press Inc.

Schütze, H. 1997. *Ambiguity Resolution in Language Learning: Computational and Cognitive Models*. Cambridge: Cambridge University Press.

Selkirk, E.O. 1983. *The Syntax of Wor*ds. Cambridge, Mass.: MIT Press.

Selting M. and Couper-KUHLEN, E. (Eds.) 2001. *Studies in Interactional Linguistics*. Amsterdam/ Philadelphia: John Benjamins.

Sharman, R. 1990. *Hidden Markov model methods for word tagging*. Report 214. Winchester: IBM UK Scientific Centre.

Sinclair, J. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smith, L. (Ed.) 1981. *English for Cross-cultural Communication*. London: Macmillan.

Souter, C. and Atwell, E. (Eds.) 1993. *Corpus Based Computational Linguistics.* Amsterdam: Rodopi.

Spencer, A. 1991. *Morphological Theory*. Oxford: Basil Blackwell.

Sperberg-Mcqueen, C.M. and Burnard, L. 1994. *Guidelines for Electronic Text Encoding and Interchange.* Chicago and Oxford: ACH-ACL-ALLC Text Encoding Initiative.

Stenström, A-B, Andersen, G. and Hasund, I.K. 2002. *Trends in Teenage Talk: Corpus Compilation, Analysis and Findings.* Amsterdam: John Benjamins.

Stern, W. 1924. *Psychology of Early Childhood up to Six Years of Age*. New York; Holt.

Stubbs, M. 1996. *Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture*. Oxford: Blackwell. Publishers.

Summers, D. 1991. *Longman/Lancaster English Language Corpus: Criteria and Design*. Harlow: Longman.

Svartvik, J. (Ed.) 1990. *The London Corpus of Spoken English: Description and Research*. Lund Studies in English 82. Lund: Lund University Press.

Svartvik, J. (Ed.) 1992. *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82 - Stockholm,* 4-8 August 1991. Berlin, New York: Mouton De Gruyter.

Tannen, D. (Ed.) 1982. *Spoken and Written Language: Exploring Orality and Literacy*. Norwood, New Jersey: Ablex Publishing Corporation.

Thomas, J. and SHORT, M. (Eds.) 1996. *Using Corpora for Language Research: Studies in the Honour of Geoffrey Leech*. London and New York: Addison Wesley Longman.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work*. Amsterdam: John Benjamins.

Uta, F. (Ed.) 1979. Cognitive Processes in Spelling. London: Academic Press.

Vera, D.E.J. (Ed.) 2002. *A Changing World of Words: Studies in English Historical Lexicography, Lexicology and Semantics.* Amsterdam: Rodopi.

Véronis, J. (Ed.) 2000. *Parallel Text Processing: Alignment and Use of Translation Corpora*. Dordrecht: Kluwer Academic Publishers.

Weigand, E. and Dascal, M. (Eds.) 2001. *Negotiation and Power in Dialogic Interaction*. Amsterdam/Philadelphia: John Benjamins.

Wichmann, A., Fligelstone, S., Mcenery, T. and Knowles, G. (Eds.) 1997. *Teaching and Language Corpora.* London: Longman.

Wills, J.D. 1990. *The Lexical Syllabus*. London: Collins.

Winograd, T. 1983. *Language as a Cognitive Process*. Vol. I. Mass.: Addison-Wesley.

Wiseby, R.A. (Ed.) 1971. *Computer in Literary and Linguistic Research*. *Papers from the Cambridge Symposium*. Cambridge: Cambridge University Press.

Young, S. and G. Bloothooft (Eds.) 1997. *Corpus-Based Methods in Language and Speech Processing.* Vol-II. Dordrecht: Kluwer Academic Press.

YULE, G.U. 1964. *The Statistical Study of Literary Vocabulary.* Cambridge: Cambridge University Press.

Zernik, U. (Ed.) 1991. *Lexical Acquisition*. Englewood Cliff, NJ: Erlbaum.

Zipf, G.K. 1936. *The Psycho-biology of Language: An Introduction to Dynamic Philology*. London: G. Routledge.

Zipf. G.K. 1949. *Human Behaviour and the Principle of Least Effort: An Introduction of Human Ecology*. Cambridge, Mass.: Addison-Wesley.